Cross-lingual Information Extraction for the Assessment and Prevention of Adverse Drug Reactions

Lisa Raithel

THESE DE DOCTORAT

NNT : 2024UPASG011

Cross-lingual Information Extraction for the Assessment and Prevention of Adverse Drug Reactions

Extraction d'information translingue pour l'évaluation et la prévention d'effets indésirables de médicaments

Thèse de doctorat de l'université Paris-Saclay et de Technische Universität Berlin

École doctorale n°580, sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : informatique
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **LISN** (Université Paris-Saclay, CNRS) et **DFKI GmbH** (Technische Universität Berlin),
sous la direction de **Pierre ZWEIGENBAUM**, Directeur de recherche CNRS,
et la co-direction de **Sebastian MÖLLER**, Professeur des universités.

Thèse soutenue à Berlin, le 09 Février 2024, par

Lisa RAITHEL

Composition du jury

Membres du jury avec voix délibérative

Matthias BÖHM Professeur des universités, Technische Universität Berlin	Président du jury
Yuki ARASE Professeure associée, Osaka University	Rapporteuse & Examinatrice
Ulf LESER Professeur des universités, Humboldt-Universität zu Berlin	Rapporteur & Examinateur
Sebastian MÖLLER Professeur des universités, Technische Universität Berlin	Rapporteur & Examinateur
Claire NÉDELLEC Directrice de recherche INRAE, Université Paris-Saclay, INRAE, MaIAGE	Examinatrice

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Extraction d'information translingue pour l'évaluation et la prévention d'effets indésirables de médicaments

Mots clés : Domaine médical, Traitement automatique des langues, médias sociaux, transfert de connaissances trans-lingue, extraction d'information, pharmacovigilance

Résumé : Les travaux décrits dans cette thèse portent sur la détection et l'extraction trans- et multilingue des effets indésirables des médicaments dans des textes biomédicaux rédigés par des non-spécialistes.

Dans un premier temps, je décris la création d'un nouveau corpus trilingue (allemand, français, japonais), centré sur l'allemand et le français, ainsi que le développement de directives, applicables à toutes les langues, pour l'annotation de contenus textuels produits par des utilisateurs de médias sociaux. Enfin, je décris le processus d'annotation et fournis un aperçu du jeu de données obtenu.

Dans un second temps, j'aborde la question de la confidentialité en matière d'utilisation de données de santé à caractère personnel. Enfin, je présente un prototype d'étude sur la façon dont les utilisateurs réagissent lorsqu'ils sont directement interrogés sur leurs expériences en matière d'effets indésirables liés à la prise de médicaments. L'étude révèle que la plupart des utilisateurs ne voient pas d'inconvénient à décrire leurs expériences quand demandé, mais que la collecte de données pourrait souffrir de la présence d'un trop grand nombre de questions.

Dans un troisième temps, j'analyse les résultats d'une potentielle seconde méthode de collecte de données sur les médias sociaux, à savoir la génération automatique de pseudo-tweets basés sur des messages Twitter réels. Dans cette analyse, je me concentre sur les défis que cette approche induit. Je conclus que de nombreuses erreurs de traduction subsistent, à la fois au niveau du sens du texte et des annotations. Je résume les leçons apprises et je présente des mesures potentielles pour améliorer les résultats.

Dans un quatrième temps, je présente des résultats expérimentaux de classification translingue de documents, en anglais et en allemand, en ce qui concerne les effets indésirables des médicaments. Pour ce faire, j'ajuste les modèles de classification sur différentes configurations de jeux de données, d'abord sur des documents anglais, puis sur des documents allemands. Je constate que l'incorporation de données d'entraînement anglaises aide à la classification de documents pertinents en allemand, mais qu'elle n'est pas suffisante pour atténuer efficacement le déséquilibre naturel des classes des documents. Néanmoins, les modèles développés semblent prometteurs et pourraient être particulièrement utiles pour collecter davantage de textes, afin d'étendre le corpus actuel et d'améliorer la détection de documents pertinents pour d'autres langues.

Dans un cinquième temps, je décris ma participation à la campagne d'évaluation n2c2 2022 de détection des médicaments qui est ensuite étendue de l'anglais à l'allemand, au français et à l'espagnol, utilisant des ensembles de données de différents sous-domaines. Je montre que le transfert trans- et multilingue fonctionne bien, mais qu'il dépend aussi fortement des types d'annotation et des définitions. Ensuite, je réutilise les modèles mentionnés précédemment pour mettre en évidence quelques résultats préliminaires sur le corpus présenté. J'observe que la détection des médicaments donne des résultats prometteurs, surtout si l'on considère que les modèles ont été ajustés sur des données d'un autre sous-domaine et appliqués sans réentraînement aux nouvelles données. En ce qui concerne la détection d'autres expressions médicales, je constate que la performance des modèles dépend fortement du type d'entité et je propose des moyens de gérer ce problème. Enfin, les travaux présentés sont résumés, et des perspectives sont discutées.

Title : Cross-lingual Information Extraction for the Assessment and Prevention of Adverse Drug Reactions

Keywords : medical domain, natural language processing, social media, cross-lingual transfer learning, information extraction, pharmacovigilance

Abstract : The work described in this thesis deals with the cross- and multi-lingual detection and extraction of adverse drug reactions in biomedical texts written by laypeople. This includes the design and creation of a multi-lingual corpus, exploring ways to collect data without harming users' privacy and investigating whether cross-lingual data can mitigate class imbalance in document classification. It further addresses the question of whether zero- and cross-lingual learning can be successful in medical entity detection across languages.

I describe the creation of a new tri-lingual corpus (German, French, Japanese) focusing on German and French, including the development of annotation guidelines applicable to any language and oriented towards user-generated texts.

I further describe the annotation process and give an overview of the resulting dataset. The data is provided with annotations on four levels : document-level, for describing if a text contains ADRs or not ; entity level for capturing relevant expressions ; attribute level to further specify these expressions ; The last level annotates relations, to extract information on how the aforementioned entities interact.

I then discuss the topic of user privacy in data about health-related issues and the question of how to collect such data for research purposes without harming the person's privacy. I provide a prototype study of how users react when they are directly asked about their experiences with ADRs. The study reveals that most people do not mind describing their experiences if asked, but that data collection might suffer from too many questions in the questionnaire.

Next, I analyze the results of a potential second way of collecting social media data : the synthetic generation of pseudo-tweets based on real Twitter messages. In the analysis, I focus on the challenges this approach entails and find, despite some preliminary cleaning, that there are still problems to be found in the translations, both with respect to the meaning of the text and the annotated labels. I therefore give anecdotal examples of what can go wrong during automatic translation, summarize the lessons learned, and present potential steps for improvements.

Subsequently, I present experimental results for cross-lingual document classification with respect to ADRs in English and German. For this, I fine-tuned classification models on different dataset configurations first on English and then on German documents, complicated by the strong label imbalance of either language's dataset. I find that incorporating English training data helps in the classification of relevant documents in German, but that it is not enough to mitigate the natural imbalance of document labels efficiently. Nevertheless, the developed models seem promising and might be particularly useful for collecting more texts describing experiences about side effects to extend the current corpus and improve the detection of relevant documents for other languages.

Next, I describe my participation in the n2c2 2022 shared task of medication detection which is then extended from English to German, French and Spanish using datasets from different sub-domains based on different annotation guidelines. I show that the multi- and cross-lingual transfer works well but also strongly depends on the annotation types and definitions. After that, I re-use the discussed models to show some preliminary results on the presented corpus, first only on medication detection and then across all the annotated entity types. I find that medication detection shows promising results, especially considering that the models were fine-tuned on data from another sub-domain and applied in a zero-shot fashion to the new data. Regarding the detection of other medical expressions, I find that the performance of the models strongly depends on the entity type and propose ways to handle this. Lastly, the presented work is summarized and future steps are discussed.

Titel : Sprachübergreifende Informationsextraktion zur Erkennung und Prävention medizinischer Nebenwirkungen

Schlüsselwörter : Medizinischer Bereich, maschinelle Sprachverarbeitung, soziale Medien, sprachübergreifender Wissenstransfer, Informationsextraktion, Pharmakovigilanz

Zusammenfassung : Die in dieser Dissertation beschriebene Arbeit befasst sich mit der mehrsprachigen Erkennung und Extraktion von unerwünschten Arzneimittelwirkungen in biomedizinischen Texten, die von Laien verfasst wurden.

Ich beschreibe die Erstellung eines neuen dreisprachigen Korpus (Deutsch, Französisch, Japanisch) mit Schwerpunkt auf Deutsch und Französisch, einschließlich der Entwicklung von Annotationsrichtlinien, die für alle Sprachen gelten und sich an nutzergenerierten Texten orientieren. Weiterhin dokumentiere ich den Annotationsprozess und gebe einen Überblick über den resultierenden Datensatz.

Anschließend gehe ich auf den Schutz der Privatsphäre der Nutzer in Bezug auf Daten über Gesundheitsprobleme ein. Ich präsentiere einen Prototyp zu einer Studie darüber, wie Nutzer reagieren, wenn sie direkt nach ihren Erfahrungen mit Nebenwirkungen befragt werden. Die Studie zeigt, dass die meisten Menschen nichts dagegen haben, ihre Erfahrungen zu schildern, wenn sie um Erlaubnis gefragt werden. Allerdings kann die Datenerhebung darunter leiden, dass der Fragebogen zu viele Fragen enthält.

Als nächstes analysiere ich die Ergebnisse einer zweiten potenziellen Methode zur Datenerhebung in sozialen Medien, der synthetischen Generierung von Pseudo-Tweets, die auf echten Twitter-Nachrichten basieren. In der Analyse konzentriere ich mich auf die Herausforderungen, die dieser Ansatz mit sich bringt, und zeige, dass trotz einer vorläufigen Bereinigung noch Probleme in den Übersetzungen zu finden sind, sowohl was die Bedeutung des Textes als auch die annotierten Tags betrifft. Ich gebe daher anekdotische Beispiele dafür, was bei einer maschinellen Übersetzung schiefgehen kann, fasse die gewonnenen Erkenntnisse zusammen und stelle potenzielle Verbesserungsmaßnahmen vor.

Weiterhin präsentiere ich experimentelle Ergebnisse für die Klassifizierung mehrsprachiger Dokumente bezüglich medizinischer Nebenwirkungen im Englischen und Deutschen. Dazu wurden Klassifikationsmodelle an verschiedenen Datensatzkonfigurationen verfeinert (fine-tuning), zunächst an englischen und dann an deutschen Dokumenten. Dieser Ansatz wurde durch das starke Ungleichgewicht der Labels in den beiden Datensätzen verkompliziert. Die Ergebnisse zeigen, dass die Einarbeitung englischer Trainingsdaten bei der Klassifizierung relevanter deutscher Dokumente hilft, aber nicht ausreicht, um das natürliche Ungleichgewicht der Dokumentenklassen wirksam abzuschwächen. Dennoch scheinen die entwickelten Modelle vielversprechend zu sein und könnten besonders nützlich sein, um weitere Texte zu sammeln. Dieser wiederum können das aktuelle Korpus erweitern und damit die Erkennung relevanter Dokumente für andere Sprachen verbessern.

Nachfolgend beschreibe ich die Teilnahme am n2c2 2022 Shared Task zur Erkennung von Medikamenten. Die Ansätze des Shared Task werden anschließend vom Englischen auf deutsche, französische und spanische Korpora ausgeweitet, indem Datensätze aus verschiedenen Teilbereichen verwendet werden, die auf unterschiedlichen Annotationsrichtlinien basieren. Ich zeige, dass die mehrsprachige Übertragung gut funktioniert, aber auch stark von den Annotationstypen und Definitionen abhängt. Im Anschluss verwende ich die besprochenen Modelle erneut, um einige vorläufige Ergebnisse für das vorgestellte Korpus zu zeigen, zunächst nur für die Erkennung von Medikamenten und dann für alle Arten von annotierten Entitäten. Die experimentellen Ergebnisse zeigen, dass die Medikamentenerkennung vielversprechende ist, insbesondere wenn man bedenkt, dass die Modelle an Daten aus einem anderen Teilbereich verfeinert und mit einem zero-shot Ansatz auf die neuen Daten angewendet wurden. In Bezug auf die Erkennung anderer medizinischer Ausdrücke stellt sich heraus, dass die Leistung der Modelle stark von der Art der Entität abhängt. Ich schlage deshalb Möglichkeiten vor, wie man dieses Problem in Zukunft angehen könnte. Abschließend werden die vorgestellten Arbeiten zusammengefasst und zukünftige Schritte diskutiert.

Cross-lingual Information Extraction for the Assessment and Prevention of Adverse Drug Reactions

vorgelegt von
M. Sc.
Lisa Raithel
ORCID: 0000-0002-5716-9566

an der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
und der
Université Paris-Saclay, Frankreich
zur Erlangung des akademischen Grades

Doktorin der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation
(Cotutelle de Thèse)

Promotionsausschuss:

Vorsitzender:
Prof. Dr. Matthias Böhm (Technische Universität Berlin)

Gutachter*innen:
Assoc.-Prof. Yuki Arase (Universität Osaka)
Claire Nédellec, Directrice de Recherche INRAE (Université Paris-Saclay)
Prof. Dr. Ulf Leser (Humboldt-Universität zu Berlin)
Prof. Dr.-Ing. Sebastian Möller (Technische Universität Berlin)

Tag der wissenschaftlichen Aussprache: 09. Februar 2024
an der Technischen Universität Berlin

Berlin 2024

Declaration of Authorship

I, Lisa Michaela Raithel, declare that this thesis titled, "Cross-lingual Information Extraction for the Assessment and Prevention of Adverse Drug Reactions" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at Université Paris-Saclay and Technische Universität Berlin.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Acknowledgements

There are so many people I would like to thank that I don't know where to start – except, of course, with my great team of supervisors! Pierre and Sebastian, thank you so much for everything you've done for me. I am very much aware that having a good Ph.D. supervisor is not something to take for granted, and I was lucky enough to have two (actually four) of them. Thank you for fighting the bureaucratic wars so this cotutelle became a reality, and for giving me this wonderful opportunity to pursue a Ph.D. in Germany, France, and Japan! I learned a lot from both of you, not only about science. Whenever I talked to either of you, I felt encouraged and reassured, especially during the writing of this thesis. Thank you for your advice, for your patience, and for your support! I hope that one day I will be as patient and meticulous in my work as you are.

Roland and Philippe, thank you for inviting me to join the team! Thanks for being my sparring partners and office mates, you were always there for me, also for last-minute proofreading, not only of this manuscript. I don't think this journey would have been the same without you and I very much appreciate the advice and support you've given me over the course of the last years. Roland, thank you for your infectious energy! Philippe, thank you for your (also infectious) calmness! Thank you both for being always ready to help.

I am very grateful to my family and all my friends, the old ones and those I met during the last three years. My sisters Julia & Lena – it is good to know that you will always be there, no matter what. My parents, who (most of the time) let me do whatever I wanted to do, for example, studying "something with language and computers", eventually resulting in this thesis.

Teele, Katja, Britta, and Ilia, who were there for me when I needed them. David, Arne, and Aleks, for the discussions, explanations, and, most importantly, the open door between our offices, the lunch and coffee breaks, and your encouragement. Hui-Syuan, my work, travel, and adventure buddy for the last (almost) three years, whose presence I'm now missing dearly. Jean, who was my tower of strength for a very long time. My friends from home, simply because we've been friends for such a long time.

And, last but not least, and in no particular order: all the lovely people from DFKI Berlin (Ibrahim, Ajay, Yuxuan, Steffen, Nils, Salar, Ekaterina, Martin, Andrés, Maria, Julián, Leo, Sven, Aljoscha, Eleftherios, Majella, and also Robert and Christoph, and many more), LISN (Mathilde, Sophie, Paritosh, Thomas, Camille, Armand, Juan, Hugo, Valentin, Mathieu, Atilla, Paul, Nesrine, Nicolas, Aurélie, Patrick, Cyril, Thomas, Agata, and many more) and the Social Computing Lab at NAIST – it was great and rewarding to work with all of you.

I also would like to thank the reviewers of my thesis: Prof. Yuki Arase, Prof. Ulf Leser, and Claire Nédéllec, Directrice de recherche. Thank you for agreeing to review and taking the time to do it!

And finally, thanks to DFG for funding my research within the KEEPHA project (DFG – 442445488) under the trilateral ANR-DFG-JST call, and to IDEX Paris-Saclay for the ADI'2020 mobility grant.

Contents

List of Abbreviations

ADR	Adverse Drug Reaction	1
CNN	Convolutional Neural Network	20
CHV	Consumer Health Vocabulary	39
CRF	Conditional Random Field	30
CUI	Concept Unique Identifier	41
DL	Deep Learning	11
EHR	Electronic Health Record	33
FP	false positive	15
FN	false negative	15
GAN	Generative Adversarial Network	32
IE	Information Extraction	3
IAA	Inter-Annotator Agreement	14
LLM	Large Language Model	42
LM	Language Model	5
LLT	Lowest Level Term	46
LDA	Latent Dirichlet Allocation	22
LSA	Latent Semantic Analysis	22
LSTM	Long Short-Term Memory Network	20
bi-LSTM	bi-directional Long Short-Term Memory Network	30
MedDRA	Medical Dictionary for Regulatory Activities	35
MeSH	Medical Subject Headings	34
ML	Machine Learning	3
MT	Machine Translation	21
MLE	Maximum Likelihood Estimation	23
MLP	Multi-Layer Perceptron	30
NER	Named Entity Recognition	4
NLP	Natural Language Processing	3
NN	Neural Network	16
PPMI	Positive Pointwise Mutual Information	22
PHI	Protected Health Information	54
POS	Part-of-Speech	35
REL	Relation Extraction	4
RNN	Recurrent Neural Network	20
SNOMED-CT	Systematized Nomenclature of Medicine–Clinical Terms	46
SVM	Support Vector Machine	16
TP	true positive	15
TN	true negative	15
UGT	user-generated text	3
UMLS	Unified Medical Language System	34
WHO	World Health Organization	129

Chapter 1

Introduction

According to Edwards and Aronson (2000), an Adverse Drug Reaction (ADR) is defined as follows:

> "*An appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product*"

These reactions can be harmful and even deadly to the patient taking a medication. According to the World Health Organization (WHO), ADRs are one of the leading causes of death around the world[1] and, as estimated in a study in Sweden, responsible for 3% of all deaths overall in the (Swedish) population (Wester et al., 2008). In a recently published study by Beeler et al. (2023), which ran over eight years in Switzerland, the authors found that 2.3% of about 32,000 hospital admissions per year were caused by ADRs.

Many of these ADRs are caused by e.g., wrong dosages, self-medication, incorrect diagnoses, or undetected conditions (like allergies) of the patient. Common reactions, as reported by Beeler et al. (2023) are, for example, hypertension, electrolyte disorders, or renal failure. Also, in general, no medication is free of side effects, and even though there are clinical trials for each drug, the pool of patients can never represent an entire population (e.g., with respect to age, gender, health, or ethnicity) (Hazell and Shakir, 2006). Even post-release surveillance campaigns might fail to reach the people who have issues with the released medication (Hazell and Shakir, 2006). Therefore, even after conducting these countermeasures, medication use and effects must be monitored constantly. For this, several public institutions for the official reporting of ADRs exist, for example spontaneous reporting systems (SRS) like the FDA Adverse Event Reporting System (FAERS) in the US, the UK Yellow Card Scheme[2] or the European Database on Adverse Drug Reaction Reports[3]. According to several studies, these reports made by patients contain more details and more explicit information about the experienced ADRs than what practitioners tend to report (Medawar et al., 2002; Herxheimer et al., 2010; Vilhelmsson, 2015).

However, even though there are official reporting authorities, ADRs suffer from serious under-reporting (Hazell and Shakir, 2006; Palleria et al., 2013). This is often due to the voluntary nature of the respective systems (Yang et al., 2012; Zolnoori et al., 2019), but even in countries where reporting serious ADRs is legally obligatory, like in Switzerland, the majority is not reported (Beeler et al., 2023). Furthermore, patients (or people taking medication) might not even know they are experiencing side effects. If they knew, they might not be aware of official places to report them (if there are any in their country) or that their physician can report them (Yang et al., 2012). Not even the physician might know that there are official contact points, and if they do, they often only report those ADRs of which they are already certain (Segura-Bedmar

[1] https://bit.ly/3tcUzoQ
[2] https://yellowcard.mhra.gov.uk/
[3] https://www.adrreports.eu/

et al., 2014). Also, not everyone likes to talk about medical problems, especially when it comes to sensitive matters, e.g., related to psychological problems or sexuality. Palleria et al. (2013) describe several other reasons for the under-reporting of ADRs. For instance, people might believe that serious ADRs are already well documented, or they are insecure if a drug is really responsible for a symptom they are experiencing. Further, they might think that their issues are not important or serious enough to be reported in the first place, they might fear consequences arising from "going public" or they might mistrust the clinical providers (Yang et al., 2012). And finally, even if ADRs are reported by professionals, the reports often lack in quality and information (Palleria et al., 2013) and technical reports often go without the information of how exactly patients suffer (Arase et al., 2020).

Consequently, other resources need to be considered for monitoring the reactions introduced by newly (and sometimes long-time) released medications. Social media presents themselves as a valuable resource since at least two thirds of the world's population have access to internet[4] and a high number thereof is active on social media[5], providing data in different languages and from different social and ethical backgrounds. Therefore, a different pool of people can be reached, and at the same time, it allows reversing the perspective to the one of the patient. This makes a big difference since non-professionals speak and write in their own voice and with their own words (even in regional dialects) about the issues they experience(d). With this, health issues that are acutely more relevant to "regular" people are exposed.

Social media, including patient fora, are a more anonymous way of sharing concerns and doubts, which might be more comfortable for patients, especially when experiencing more tabooed side effects. Depending on the used platform, users might not need to disclose their names or any other personal information. They can communicate their concerns openly and without fear of consequences. Of course, this also makes it more difficult to verify any information they share, touching the topic of factuality and fake news in times of worldwide communication.

Another factor for turning to social media is, as already mentioned, the variety of languages provided on the internet. Most scientific publications are written and distributed in English to reach a broader research community. However, most (lay) people do not understand these publications – either because they speak a different language than English, because it is written in a very technical language or both. Therefore, they turn to the World Wide Web, often patient fora, to research and collect information on topics they are concerned with, following "translations" from technical terminology to layperson language provided by other members of the respective communities. Sometimes, there are even clinicians involved in these fora. By now, several initiatives[6] in various countries exist with the goal of providing health information in a way that is understandable for laypeople. This, again, highlights the necessity to extract relevant information not only from texts written by experts but also to listen to the patients' voices and process texts written by "normal" people. In the long term, it also might help clinicians and other practitioners to understand their patients and the experienced ADRs better, react more appropriately, and meet the patients' needs more precisely (Arase et al., 2020), allowing the patients to participate actively in their own treatment (Segura-Bedmar et al., 2014). Finally, the collected information from "crowd signals" (Scaboro et al., 2022) can be used for drug re-purposing and the development of new medications[7].

Note that when using data from social media, we again commit to a certain sub-group of people: Those who have access to and actively participate on these platforms. Depending on the platform, the age range might vary, too.

[4] https://www.statista.com/topics/1145/internet-usage-worldwide

[5] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[6] For example, https://www.patienten-information.de/leichte-sprache or https://washabich.de/.

[7] Of course, all detected information should be reviewed by at least one recognized expert in the field.

The previous paragraph established that multi-lingual user-generated texts (UGTs) collected from social media are a valuable resource that should definitely be used to support pharmacovigilance. Now we need the means to do this efficiently. No human can possibly go through the huge amounts of text published every second on the internet. Yet, being able to process and present data quickly allows to extract information in a concise and structured way and to react immediately. In the case of pharmacovigilance, it can trigger medical investigations (Scaboro et al., 2022) which might be a matter of life and death in the worst case (Beeler et al., 2023). However, texts collected online are usually unstructured, not standardized, and contain much more content than is needed for pharmacovigilance. For demonstration purposes, see Example 1.1, a drug review collected for the corpus PSYTAR (Zolnoori et al., 2019) from the forum AskAPatient[8]. Texts like the one displayed need to be processed and filtered, and the relevant information needs to be extracted.

(1.1) *"Wow this stuff is strong, i took the first 10mg today, and can't focus, i'm totally apathetic. This lexapro works too much like a neuroleptic. i hate feeling out of control, detached. I would rather be depressed than totally anihilated ! Never again."*[9]

This is exactly where methods from Computational Linguistics or Natural Language Processing (NLP) are useful. NLP is the application of computational methods for the analysis and synthesis of natural text and speech. In particular, the sub-field of Information Extraction (IE), a type of information retrieval, is of interest in this case. It is dedicated to automatically acquiring information relevant to a specific task or question by collecting it from texts and presenting it in a structured way, e.g., in a database (Sarawagi, 2008). For example, commonly known applications are sentiment analysis, i.e., attributing a product review with a certain sentiment (positive, negative, neutral), or event extraction, e.g., extracting information relevant to an event (date, time, location) and the event itself from an e-mail.

In the presented use case, we are interested in drug mentions and medical signs and symptoms, and the relationships between those, amongst other things. For example, from a document like the one in Example 1.1, we would like to extract the information that the patient took the medication *Lexapro* because of a certain diagnosis (*I would rather be depressed*) and that they experience side effects like `impaired concentration` (*can't focus*), and `apathy` (*totally apathetic*) and a `feeling of helplessness` and `detachment` (*feeling out of control*, *feeling detached*, *being anihilated*). Furthermore, the patient gives us information about the drug's dosage (*10mg*) and implicitly says that, although they just started (*took the first 10mg today*), they will stop the medication immediately (*Never again.*), supposedly because of the adverse reactions.

Nowadays, automatically extracting this kind of information is typically done by using (neural) Machine Learning (ML) approaches. Often, *labeled* data is used to *train* a model and then apply this model to relevant data, e.g., new incoming social media texts, to extract information similar to the one that was labeled in the training data. Usually, several steps are required to successfully set up a pipeline that gets as input a plain document collected from the internet and returns structured information that can be re-used for further tasks. We now assume that we want to build a ML model that does exactly that. The necessary steps are briefly described as follows:

Data Collection First of all, data for a model to *learn* and to be *evaluated* on needs to be collected. The difficulty of that depends very much on the domain and language in which the data is supposed to be collected and the task it is required to help in. For example, getting English product reviews might be easier than collecting Italian legal texts. Also, accessibility, copyrights, and (user) privacy might play a role in the collection procedure.

[8] `https://www.askapatient.com/`

[9] Example from PsyTAR corpus (Zolnoori et al., 2019).

Data Annotation Labeling or *annotating* is the process of adding analytical or descriptive notes to a stream of raw text. For instance, in the text provided in Example 1.1, we would mark the medication name *"lexapro"*, which is an *entity* we are interested in, with a label named `drug` and *10mg* with the label `dosage`.

Data (Pre-)processing Pre-processing might happen before and after data annotation. It includes, for example, cleaning and tokenizing the data. Cleaning often involves removing noise or other disturbing factors, e.g., hashtags from social media. Tokenization is the process of separating a stream of data (text) into smaller, meaningful parts, usually paragraphs, sentences, and single tokens[10].

At this point, data *de-identification* might also be applied, a pre-processing step that is particularly relevant in the biomedical or clinical domain. This involves methods that obscure or remove mentions of personally identifying information that might make it possible to back-trace the text to its original author.

Named Entity Recognition (NER) After data preparation, the extraction of information can begin. NER is the task of extracting entities or spans from a given text and classifying them into pre-defined labels. Entities are mostly self-contained expressions, like locations or drug mentions (e.g., *"lexapro"* as `drug`) while spans can comprise longer sentence fragments, e.g., colloquial disorder descriptions (e.g., *"can't focus"* as `disorder`). Given a text, we want to receive the entities of interest tagged with their corresponding labels, but also the position of these entities in the text. In the case of ADRs the syntactic position of the drug mentioned might be important to help in judging if a symptom is the *reason* or *outcome* of the drug. See Figure 1.1 for an example.

Relation Extraction (REL) This is the task of identifying and classifying relationships between the aforementioned entities or spans. For instance, in the presented example, there exists a relationship between the mention *lexapro* with the label `drug` and the mention *can't focus* with the label `disorder`. The relationship might be called `caused` in this case since the medication seems to be the reason for the symptom. Again, please refer to Figure 1.1 for a visualization.

Entity Linking / Entity Normalization This, finally, is the task of associating an entity with its normalized version, which usually comes from an ontology or taxonomy. It means to unify, and ground mentions from unstructured documents to a reference concept or category to be further processed.[11]

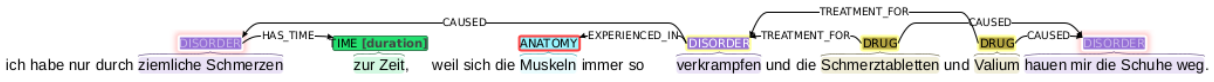


Figure 1.1: An example annotation of a German text from the newly created corpus. Entities as well as relations are annotated, additional attributes (`duration` attached to the `time` expression) provide more fine-grained information.

We will re-visit all of the listed tasks in the remainder of this thesis except the step of entity normalization since this is ongoing and future work. However, it still illustrates how extracted information can be further processed and made available to physicians or other end-users.

While there exist established and decently working methods for all described steps, these are not easily transferable to (i) the biomedical domain, (ii) social media data and (iii) languages other than English. For example, regarding (i), we often are confronted with a skewed

[10] A token, as defined by Manning et al. (2009), is "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.".

[11] Entity Normalization is not discussed in this thesis; it is added for completeness.

data distribution in the medical domain. Events or documents relevant to the targeted use case are not very frequent compared to other occurrences, but they are important nonetheless. In the case of ADRs, documents containing relevant descriptions of experiences are in the minority if the forum is not dedicated to this exact topic. Also, the biomedical domain has its own jargon, including abbreviations which can vary from hospital to hospital.

As for (ii), it is often difficult to deal with social media data in general since people use a very different language to what is usually considered "correct" in terms of grammar, orthography, or lexis. This is particularly applicable when addressing medical topics. Here, patients might know nothing about their symptoms and simply describe their feelings, i.e., use colloquial everyday language. Still, they might also be experts with respect to a certain disease and therefore use the exact same terminology as practitioners. Additionally, the use of abbreviations (medical or colloquial) and emojis can complicate processing. This provides a very interesting field to study but makes it much harder for ML models to generalize.

Finally, coming to (iii), and summing up the above, these challenges are already difficult for English but even harder for other languages. This is mostly due to the scarce data available in this domain and languages other than English, often caused by privacy issues and a smaller NLP community to support the creation of datasets. It also extends into the lower availability of pre-trained language- or domain-specific Language Models (LMs)[12], although there has been improvement in recent years. Nevertheless, even if data are easy to come by in certain circumstances or for specific languages, they still need to be annotated to be used as training material. These annotations are based on guidelines and schemes, usually defined for one language and an exact purpose. This makes it intricate to transfer them to other, even related, domains or languages for which, in turn, new datasets, guidelines, and schemes have to be developed.

Based on the above-mentioned issues in the current research with respect to the cross-lingual detection of Adverse Drug Reactions, we infer the following research questions, which are attempted to be answered in the remainder of this thesis.

Research Questions & Contributions

This work describes the following Research Questions (RQs) and contributions to tackle the tasks of classifying and extracting Adverse Drug Reactions from user-generated texts within and across languages.

Research Questions

Research Question 1. *Can we create annotations with respect to ADRs across languages and how do annotation guidelines need to be designed such that they are applicable to all targeted languages?*

Research Question 2. *Is it possible to collect high-quality descriptions of medical side effects by simply asking patients online to describe their experiences?*

Research Question 3. *Can few-shot approaches improve classification performance when faced with a high label imbalance?*

Research Question 4. *Do multi-lingual transformer models work well enough to reliably extract drug mentions from texts originating from different genres in different languages?*

Research Question 5. *How well does the transfer of learned knowledge about ADRs from one language to the other work in the bio-medical domain?*

[12] Generative (large language) models are not discussed in this thesis.

Contributions

Many of the following contributions were a joint effort within KEEPHA[13], a trilateral project between Germany, Japan, and France.

- Development of *cross-lingual* annotation guidelines for user-generated content in the bio-medical domain. The guidelines are applicable to (at least) four languages: English, German, French and Japanese. Note that these languages represent three different language families and that their speakers come from diverse cultural backgrounds. This contribution refers to RQ 1 and is discussed in Section 4.1.
  The general guidelines were developed in a group effort of team KEEPHA. I was responsible for developing and validating the guidelines based on German examples, instructing and training annotators and supervising the annotation process. Problems and questions or unclear instructions regarding annotation were discussed between the annotators and me, then brought to monthly meetings to discuss with the team and consolidate with the other languages. I was further responsible for consolidating the German data, both the binary and entity/relation annotations. The French data was prepared by me as well. Similarly to the German data, the annotators' questions and problems were discussed in weekly meetings and whenever needed.

- Creation of a corpus of 118 annotated documents in German containing entity, attribute, and relation-level annotations according to the guidelines, the first of this kind for research on the extraction of ADRs. This contribution refers to RQ 1 and is discussed in Section 4.1.
  I was responsible for aggregating and preparing the data, as well as training the annotators for both the German and the French datasets based on the above mentioned guidelines. I curated the German data, calculated inter-annotator agreement and analyzed the results.

- Creation of a corpus of 10,000 documents in German and 864 documents in French with binary annotations for document-level classification of ADRs. This contribution refers to RQ 1 and is discussed in Section 4.1.
  I collected the data, set up the annotation system, trained the annotators and discussed annotations with them. I further curated the dataset, calculated inter-annotator agreement and analyzed the results. I further supervised the annotators when checking the translations into French.

- A prototype study for gathering data relevant to bio-medical text processing so that patients can consent to the processing. This contribution refers to RQ 2 and is discussed in Section 4.2.1.
  This study was a result of the Usable Privacy seminar at TU Berlin. I proposed the idea, designed the questionnaire and ran the study. Subsequently, I analyzed the results.

- Experiments on the binary classification of German documents containing ADRs in a low-resource setting, using zero- and few-shot techniques and cross-lingual knowledge transfer from English to German. The high label imbalance in both languages is a big challenge and few-shot approaches, even when using balanced label distributions, are less helpful than fine-tuning on the highly imbalanced original dataset. This contribution refers to RQ 3 and is discussed in Chapter 5.
  *This was published in **Raithel** et al. (2022); For detailed contributions, see below.*

[13]Participating project partners are TU Berlin, DFKI GmbH Berlin (Germany), Riken, NII, and NAIST (Japan), and LISN, CNRS, Université Paris-Saclay (France); https://keepha.lisn.upsaclay.fr/wiki/doku.php?id=start, within the ANR-DFG-JST trilateral call on Artificial Intelligence.

- An analysis of the performance of multi-lingual language models for drug detection across languages (German, French, Spanish, English) and within language sets (German/English & French/Spanish). We find that the performance of the multi-lingual models strongly depends on the underlying annotations; even within languages, knowledge sometimes fails to be transferred. This contribution refers to RQ 4 and is discussed in Section 6.2.
  This work was done jointly with Johann Frei, Augsburg University, supervised by Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Sebastian Möller, and Frank Kramer. JF prepared and analyzed the data and both JF and I designed and discussed the experiments. I set up and ran the experiments, aggregated the results, and analyzed the models' errors. Both JF and I wrote the manuscript, supervised and reviewed by PT, RR, PZ, SM, and FK.

- A first NER baseline for the newly created KEEPHA dataset. The baseline already shows promising results, within and across languages, but a high variation depending on entity type. This contribution refers to RQ 5 and is discussed in Section 6.3.
  I prepared, ran and analyzed the experiments.

- Support in preparing and validating a parallel multi-lingual dataset containing synthetically created tweets in four languages for the NTCIR-17 social media shared task. This contribution is discussed in Section 4.2.2.
  Joint work with the KEEPHA team for NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection Subtask; See detailed contributions below.

Publications

The research of this thesis has so far resulted in the following publications:

1. **Lisa Raithel**, Philippe Thomas, Roland Roller, Oliver Sapina, Sebastian Möller, and Pierre Zweigenbaum. 2022. Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3637–3649, Marseille. European Language Resources Association
   ***LR** collected and prepared the data, set up the annotation system, and created the (binary) annotation guidelines. OS annotated most of the data, supported by **LR**. Problems and questions during annotation were discussed between OS and **LR**. Together with PZ, **LR** designed the experiments. **LR** implemented the experiments, evaluated the performance and analyzed the results. **LR** wrote the manuscript, which PT, RR, SM, and PZ then reviewed. Note that we only published approximately half of the binary annotated data in this paper.*

2. **Lisa Raithel***, Faith W. Mutinda*, Gabriel H. B. Andrade*, Hui-Syuan Yeh*, Tomohiro Nishiyama*, Mathieu Laï-King, Shuntaro Yada, Roland Roller, Cyril Grouin, Agata Savary, Aurélie Névéol, Thomas Lavergne, Eiji Aramaki, Sebastian Möller, Yuji Matsumoto, and Pierre Zweigenbaum: KEEPHA at n2c2 2022: Track 1 Contextualized Medication Event Extraction, 2022. *n2c2 Shared Task and Workshop, 2022* (2022/11/4, Washington, D.C.) (no proceedings)

   *HY and **LR** proposed participating in the shared task. **LR** implemented and evaluated the experiments on medication extraction (subtask 1), FM and GA implemented and evaluated the experiments on event classification (subtask 2), and HY implemented and evaluated the experiments on context classification (subtask 3). **LR** conducted a detailed error analysis for subtask 1. **LR** coordinated meetings across subtasks and subtask groups and drafted the technical report. **LR** further coordinated the preparation for the talk the team was invited to and presented it with TN and FM.*

All other authors contributed to the meetings and reviewed the technical report.
This contribution is discussed in Section 6.1.

Publications that were in preparation or under review when writing this manuscript and will be referred to in this thesis:

1. **Lisa Raithel***, Hui-Syuan Yeh*, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Philippe Thomas, Sebastian Möller, Tomohiro Nishiyama, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum. 2024. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages. In *Proceedings of the Language Resources and Evaluation Conference*, Torino. European Language Resources Association
This paper describes the KEEPHA corpus, also discussed in this thesis, and accompanying experiments. ***LR*** *was responsible for the French and German parts of the tri-lingual corpus, collected and prepared the data, and was a main contributor to the annotation guidelines together with SY.* ***LR*** *supervised the annotation of the data and curated the German part. Problems and specific phenomena in the annotation process were first discussed between the annotators and* ***LR****, and if necessary, with all authors. HS and* ***LR*** *discussed, designed, and evaluated the experiments, most of which were conducted by HS. HS and* ***LR*** *wrote the first draft of the manuscript. The final paper was written with the help of all other authors.*

2. Shoko Wakamiya*, Lis Kanashiro Pereira*, **Lisa Raithel***, Hui-Syuan Yeh*, Peitao Han*, Seiji Shimizu*, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas*, Roland Roller*, Pierre Zweigenbaum*, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada†, Eiji Aramaki†. 2023. *NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview.*
EA, SW, and SY proposed this shared task. TN, GA, and SS produced the initial version of the corpus, including the translations. NN, HT, NT, YM, AA, SM, TL, and PP discussed the corpus design. PZ and PT developed the label validation methods. ***LR*** *and HY developed the evaluation scripts.* ***LR*** *coordinated the exchange across teams and communicated the results of the automatic translation validation to the Japanese team.* ***LR*** *further reviewed many German and some French translations and labels. RR, PT, AN, CG, and PZ discussed the corpus and label design and helped with the multilingual support. PH and LP built the baseline systems and evaluated the results.*

Other publications that are not directly referred to in this thesis are as follows:

1. Roland Roller, Ammer Ayach, and **Lisa Raithel**. 2021. Boosting transformers using background knowledge, or how to detect drug mentions in social media using limited data. In *Proceedings of the BioCreative VII Challenge Evaluation Workshop*
I helped with some experiments and the writing of the manuscript.

Organization of the thesis The thesis is organized in the following chapters: After the introduction and motivation of this work in Chapter 1, Chapter 2 presents the theoretical and conceptual background of the described work. Next, Chapter 3 sets the content of this thesis into the context of related work in cross- and multi-lingual information extraction in general (Section 3.1), information extraction in biomedical natural language processing (Section 3.2) and the combination of those two, cross- and multi-lingual information extraction in the domain of biomedical natural language processing (Section 3.3). The manuscript is then further divided into four themed chapters: Section 4.1 presents the annotation guideline development

* equal contribution, † equal leadership

and corpus creation process. A brief detour to data privacy in the biomedical domain follows in Section 4.2, which is in turn followed by an analysis of the problems introduced when generating synthetic tweets and translating these, as done for the NTCIR-17 shared task. Chapter 5 describes the next step in detecting ADRs: the classification of documents into those containing mentions of adverse reactions versus those that do not. Following that, cross- and multi-lingual entity extraction in the biomedical domain is discussed in Chapter 6, focusing on the detection of drug names in English data (Section 6.1) and cross-lingual detection of medication mentions in Section 6.2. The chapter concludes with preliminary experiments on the newly developed dataset in Section 6.3. Finally, in Chapter 7, the work described in this thesis is concluded and possible future research directions are outlined. Technical details and additional results and information can be found in the appendices.

Chapter 2

Background

Extracting information from natural, unstructured texts is a long and widely studied topic in NLP. As already mentioned, this includes, for instance, document classification, Named Entity Recognition, Relation Extraction, part-of-speech tagging, dependency parsing, or anaphora resolution. While rule- and regular expression-based approaches were common some years ago (and are still in use today), the advent of ML and especially Deep Learning (DL) increased the development of new methods for extracting relevant information by far.

The following will provide the background knowledge for the rest of the thesis. First, the definitions of the tackled tasks are given in Section 2.1, completed with the respective evaluation methods in Section 2.2. We will then review the needed concepts in Machine Learning (Section 2.3): Basic algorithms and learning techniques, including the concept of transfer learning (Section 2.3.4) and models commonly used for the tasks described in this manuscript (Section 2.3.5). The techniques of language modeling are addressed in Section 2.4, with a particular focus on Transformer-based models in Section 2.4.4 and Section 2.4.5.

2.1 Task Definitions

Document Classification Document classification or categorization aims to assign one or more labels to a given document. Documents can be basically anything, spanning, for instance, texts, images, and videos. In this thesis, "document" always refers to text that might be represented as, for instance, a sentence, a paragraph, or an entire user post on social media.

The labels in document classification are usually pre-defined. Common tasks are, for example, sentiment classification with the labels *positive*, *negative*, *neutral*, or news classification, i.e., categorizing news articles into their main topics, e.g., *politics*, *science*, or *sports*.

Named Entity Recognition (NER) Initially called "Named entity recognition and classification (NERC)", this task is focused on the detection and classification of (mostly) proper nouns, i.e., words or phrases denoting person names, locations, organizations, or other individual *named entities*. Over the years, the term "named entity" in the context of IE was relaxed and started to include other types, like temporal or numerical expressions. Also, the types became more fine-grained, e.g., "location" can be split into sub-types such as "country", "state", "city" (Nadeau and Sekine, 2007). Since the expressions that need to be extracted can also span more than one word, NER is also often referred to as *span detection* (and classification).

As the name suggests, the task consists of two parts: First, the span of interest has to be detected and its boundaries need to be found. Second, the extracted span needs to be classified into one out of a set of pre-defined semantic *types*. The task is usually modeled as a sequence labeling task: A sequence of tokens is given to the system, which then returns a sequence of types in the same length. For this, the problem is converted into the task of *token classification*. That implies that the sequence first needs to be *tokenized*, i.e., split into lexical units. A very simple version of this is to split the sequence by spaces. Then, for NER, every token gets

assigned an entity type (or *tag*). An often employed tagging scheme is the `BIO` scheme (also `IOB`) (Ramshaw and Marcus, 1995), representing the **B**eginning, **I**nside and **O**utside of an entity. These tags can also have more detailed attributes to distinguish between different entity types, for instance, `B-Country` or `B-State` to mark tokens belonging to a country or state expression, respectively.

Relation Extraction (REL) This task involves extracting semantic relations between entities or spans from unstructured texts. Usually, entity spans and types are given (except in joint NER and REL systems), as well as a pre-defined set of relations. Therefore, given two entities, the relation between them only needs to be classified. This task is also often called *relation classification*.

Note that the *direction* of the relation also matters, depending on the task. Moreover, the same entities can stand in different relationships, subject to the context. For example, in Figure 1.1, the context of the relationship between a medication mention and a disorder decides if the medication is a `TREATMENT` for a disorder, or if the disorder was `CAUSED` by the medication. Finally, in theory, relations can also have more than two arguments. However, in this thesis, we are only concerned with binary relation extraction, i.e., each relation has exactly two arguments, the *head* and *tail*. REL can be used, for instance, to classify drug-drug interactions or relations between persons and companies in financial news.[1]

2.2 Evaluation of Predictions and Annotations

Both predictions made by systems but also annotations (usually) made by humans need to be evaluated and validated. Standard methods are discussed in the next two sections.

2.2.1 Task Evaluation

This section briefly explains how the above-described tasks are evaluated. Usually, the systems' predictions are evaluated against a *gold standard* dataset, i.e., a dataset containing curated annotations, e.g., marked entities and their types (classes). Usually, *precision*, *recall* and F_β *score* are used. We use (binary) document classification as a running example to explain these scores. A *positive* document is relevant for us and *negative* documents are all other documents in which we are not interested. Precision and recall are then defined using the following counts:

TPs: True positives, the number of correct hits with respect to positive documents.

TNs: True negatives, the number of correct hits with respect to negative documents.

FPs: False positives, the number of negative documents incorrectly labeled as positive.

FNs: False negatives, the number of positive documents incorrectly labeled as negative.

Precision, then, is the proportion of true positives out of all documents predicted as positive (Equation 2.1).

$$precision = \frac{TPs}{TPs + FPs} \tag{2.1}$$

Recall, on the other hand, is the proportion of the true positives against the sum of all possible (correct) positives in the dataset (Equation 2.2).

[1] Although the task of REL is not performed in this thesis, we annotate the later described dataset with relations as well and therefore mention the task here.

$$recall = \frac{TPs}{TPs + FNs} \tag{2.2}$$

There is always a trade-off between those two measures and systems are often optimized for one or the other, depending on the task. Let's take document classification as an example again. If we are interested in a specific type of document but assume that there might not be that many relevant documents overall, we might be more interested in recall: A high recall tells us that we retrieved most of the relevant documents, even if some of them might not be actually positive in the end. A recall of 1.0 would mean that we retrieved *all* relevant documents (and maybe other non-relevant ones). In case we are more interested in the documents predicted as positive to be actually positive and do not mind missing some other relevant ones, we should optimize for precision. A precision of 1.0 would mean that all the documents we retrieved are positive. However, there might be other positive documents that were not retrieved.

To combine precision and recall, the commonly used measure in IE is F_β score, as shown in Equation 2.3 and originally introduced by Rijsbergen (1979). It is usually used in the form of the harmonic mean between precision and recall, i.e., with $\beta = 1$ (Chinchor, 1992). Setting β to a lower or higher value emphasizes precision or recall, respectively.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{2.3}$$

All three measures have a range between 0 and 1; the higher the resulting score, the better the system's performance. Calculating F_β score can also be applied to multi-class and multi-label settings[2], then it depends on how the per-class scores are averaged over the classes. Mostly used are *micro*, *macro* and *weighted* averages, defined in the *sklearn* library[3] (Pedregosa et al., 2011) as follows:

micro average: Calculates F_β by counting TPs, FPs, and FNs globally, no matter the class, and does therefore not take label imbalance into account.

macro average: Calculates F_β score per class and returns the unweighted mean, i.e., each class is treated equally, no matter the number of samples.

weighted average: Calculates F_β score per class and weighs per-class scores by *support*, i.e., the respective number of classes in the test set. This accounts for class imbalance.

Another widely used measure is *accuracy* (Equation 2.4):

$$acc = \frac{TPs + TN}{TPs + TNs + FPs + FNs} \tag{2.4}$$

However, this measure can be massively misleading when used on imbalanced data, obscuring the system's actual performance for document classes with a low representation. Of course, F_β score and its different averages have their biases, too, and should not be used in certain cases (Manning, 2006; Powers, 2020; Harbecke et al., 2022). The above metrics were all described using the example of (binary) document classification. NER and REL evaluation measures are also based on them, but some peculiarities need to be considered.

[2]Multi-class: There are more than two classes available for classification, but a document can only be assigned one class at a time. Multi-label: There are more than two classes available and every document can be assigned to more than one class.

[3]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Evaluation of Named Entity Recognition & Relation Extraction

For NER, two aspects need to be taken into account for evaluation: First, the predicted span has to match with the gold standard span. Second, the type predicted for the span has to match the one given in the gold data. There are also ways to analyze NER in more detail, categorizing potential errors in different classes, as presented by Chinchor and Sundheim (1993).For the NER systems presented in this thesis, we always evaluate the BRAT format of the predictions, where both span offsets and entity types are considered.

The evaluation of REL works similarly. We either assume that the entities are already given and, therefore, only evaluate the relations between these, like in a multi-class classification scenario. In case the entities are not given (joint NER and REL), then errors made in offsets or entity type prediction are propagated to the relation classification. Therefore, we also need to consider if the relation was classified correctly, adding another "layer" on top of the metrics described above.

The above approach, only counting *exact* matches as correct, is also often called *strict* evaluation. However, a more *relaxed* approach, also denoted *lenient* matching, is to consider entity spans as correct matches as long as they overlap to a certain degree. This might depend on a threshold with respect to the number of characters or a percentage of the tokens that must overlap with the gold span. For the evaluation of entity spans in this work, we use the BRAT format evaluation script[4] provided at the CLEF eHealth 2015 Task 1b (Névéol et al., 2015), originally developed by Verspoor et al. (2013). For calculating the number of lenient matches, we use the option `overlap`, which counts matches as correct as long as there is any kind of overlap between the predicted and gold span.

2.2.2 Annotation Evaluation

Inter-Annotator Agreement (IAA) is a measure to quantify the consensus of different annotators on the same data. As we will see when reviewing the existing datasets for the targeted domain, several annotation evaluation metrics are possible. Since we annotated our data on several levels (document-wise binary annotation, entities, attributes, and relations), we are in need of appropriate measures for each level. In the following, we briefly describe some of the possible measures and the ones we chose to evaluate our data's annotation quality.

		a2	
		positive	negative
a1	positive	*a*	*b*
	negative	*c*	*d*

Table 2.1: A comparison between annotator a1 and annotator a2. *a*: number of times both annotators agree on the positive label, *d*: both annotators agree on the negative label, *b*: annotator a1 rates positive, a2 rates negative, *c*: a1 rates negative, a2 rates positive.

Consider Table 2.1, which shows the number of times each annotator a1 and a2 agreed on the positive label (a) and the negative label (d), and the number of times a1 chose the positive label while a2 chose the negative label (b) and a1 chose the negative labels while a2 chose the positive label (c). The most basic form of an agreement measure is achieved by simply calculating the percentage of labels that all annotators agreed on (*observed agreement* $A_{observed}$, Equation 2.5) (Hripcsak and Heitjan, 2002), equivalent to the accuracy measure in Equation 2.4.

$$A_{observed} = \frac{a+d}{a+b+c+d} \tag{2.5}$$

[4] https://perso.limsi.fr/pz/blah2015/

However, for each annotation task, there is also a chance that annotators coincidentally chose the same label, which is not represented by this score. It further does not account for different (dis-)agreements of various categories, i.e., annotators might agree on one category more than on another (Hripcsak and Heitjan, 2002; Hripcsak and Rothschild, 2005). The latter is resolved by a measure called *specific agreement* (Cicchetti and Feinstein, 1990), which aims to calculate agreement for each category, i.e., the positive (A_{pos}) and negative (A_{neg}) category in this example.

$$A_{pos} = \frac{2a}{2a+b+c}; \qquad A_{neg} = \frac{2d}{b+c+2d} \tag{2.6}$$

To account for chance agreement, a further metric was introduced: The κ score, often called Fleiss' κ (Fleiss, 1975) or Cohen's κ (Cohen, 1960), with the difference that the version provided by Fleiss (1975) allows for more than two raters. In general, κ is based on observed agreement, as defined in Equation 2.5, and *expected agreement* $A_{expected}$ introduced by chance, as defined in Equation 2.7.

$$A_{expected} = \frac{E[a] + E[d]}{a+b+c+d}, \tag{2.7}$$

where $E[a] = \frac{(a+b)(a+c)}{a+b+c+d}$ and $E[d] = \frac{(b+d)(c+d)}{a+b+c+d}$, the expected values for a and d. Equation 2.8, thus, reduces chance agreement to zero.

$$\kappa_{Cohen} = \frac{A_{observed} - A_{expected}}{1 - A_{expected}} \tag{2.8}$$

For both the observed agreement and the κ score, the number of negative samples (d) needs to be known, which is not always the case. Moreover, if the number of negative samples is known to be large, then it is unlikely that annotators will agree on positive samples only by chance, that is, the probability of agreement by chance on positive samples is close to zero (Hripcsak and Rothschild, 2005). This is indeed often the case for tasks such as NER or REL: First of all, the notion of "negative examples" is ill-defined in some cases (Hripcsak and Rothschild, 2005). Second, if all no-entity annotations are considered, i.e., all tokens not marked as any kind of entity, this might be a huge number. According to Hripcsak and Rothschild (2005), in case of a very large number of negative samples, Equation 2.8 approaches Equation 2.6, i.e., the κ score approaches positive agreement. Although κ is widely used in information retrieval, there is also no consensus on which score reliably represents a strong agreement between annotators, which also depends on the task.

Finally, the F_1 score, a metric already discussed for the evaluation of NER and REL system predictions, can be considered as well for evaluating IAA. Here, the annotations of one annotator are handled as gold standard data, while the annotations of the second rater are treated as system predictions. Precision and recall can then be calculated as shown in Equation 2.9.

$$precision = \frac{a}{a+b}; \qquad recall = \frac{a}{a+c} \tag{2.9}$$

Note that a is equivalent to true positives (TPs), b is equivalent to false positives (FPs), and c is equivalent to false negative (FN). For neither score, we need to know the number of true negatives (TNs) (d). F_1 score is then formulated as shown in Equation 2.10 (Hripcsak and Rothschild, 2005), which is equivalent to Equation 2.3. Then, the balanced F_1 score is also equivalent to positive specific agreement as given in Equation 2.6.

$$F_1 = \frac{2a}{2a+b+c} \tag{2.10}$$

There are also various other metrics, like Scott's π (Scott, 1955), a correlation statistic that only differs to Cohen's κ in the way how the expected values $E[a]$ and $E[d]$ are calculated. Finally, Krippendorff's α (Krippendorff, 2004) is a more robust version of Cohen's κ which can be calculated with any number of annotators and categories. All these correlation measures have the problem that they are difficult to interpret (Hripcsak and Rothschild, 2005).

Annotation is a subjective process that can include biases. Therefore, there is no absolute "truth" to rely on (Grouin et al., 2011), it is even possible that all annotators are correct, even if they disagree with each other. We can only evaluate to which extent the annotators are consistent with each other, not if they annotated correctly.

As mentioned, Cohen's κ is problematic particularly for tasks with an unknown amount of negative samples, while observed agreement does not take chance into account. We therefore follow Grouin et al. (2011) and use F_1 score for entity, relation and attribute annotation evaluation, where we do not know the number of negative samples, whereas we use Cohen's κ as well as F_1 score for the evaluation of our binary annotations, where we do know the number of negative samples, but also observe a high imbalance in the label distribution.

2.3 Machine Learning

In this section, we will briefly explain the foundations of this thesis: Machine Learning (ML) and Neural Networks (NNs). Most of the theoretical part is taken from the popular *Speech and Language Technology Processing* book by Jurafsky and Martin (2023) if not otherwise stated.

An often cited early definition of Machine Learning is attributed to Mitchell (1997):

Definition 1 (Machine Learning)**.** *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*

In NLP, and specifically in this thesis, the task T is any of the ones presented: document classification, NER, or REL. P might be one (or all) of the measures described in Section 2.2. The experience E is, usually, a large collection of data points or observations, i.e., what is called a *corpus* in (computational) linguistics. Available datasets are often split into three parts: A *training* set (train) which serves as material to learn from (the experience E), a *development* set (dev), and a *test* set. On the development set, the model is evaluated *during* training; on the test set, it is evaluated *after* training to check performance on completely unseen data.

Machine Learning attempts to model data mathematically with the goal of predicting or generating a certain output for a given input. This procedure is often called *learning*: A model learns patterns from a given dataset and then applies these patterns to new, unseen data. The model itself is a function f that tries to represent the distribution of the underlying data as accurately as possible[5]. The data points (examples) are represented as feature vectors to the model.

Types of models or algorithms used in "traditional" ML are, for example, Naive Bayes or Support Vector Machines (SVMs). They are mostly based on hand-crafted features, e.g., vectors representing the occurrence of specific keywords in a given text, the capitalization of tokens, or the syntactic type of a token. A very common and powerful feature are word or sentence embeddings, which will be described in Section 2.4.

[5]Note that if a model fits the data too well, it might be overfitting and will not generalize to other datasets.

2.3.1 Deep Machine Learning

Deep Machine Learning refers to the use of artificial Neural Networks (NNs) in combination with ML. An NN consists of many computing units (functions) that take in a vector of input values and output a single value. See Equation 2.11 for a representation of one computing unit.

$$z = b + \sum_i w_i x_i \tag{2.11}$$

A unit is taking a weighted sum z of its input vector values $x_1, ..., x_n$ and the associated (learned) weights $w_1, ..., w_n$. The bias term b is added. Finally, instead of using z directly to propagate further, a neural unit applies a non-linear *activation function* f to z. Popular activation functions are, for instance, *sigmoid*, *tanh* or the rectified linear unit (*ReLU*). The result of this is commonly called the *activation value* a of z. If this is the model's final output, it is called y.

$$y = a = f(z) \tag{2.12}$$

Because of the combination of many (non-linear) functions, NNs are capable of modeling more complex representations of data than "traditional" ML approaches. One of the simplest NNs is a feed-forward network. The name stems from the fact that input information (features) is propagated iteratively from one layer made of computing units to the next without cycles. Thus, the output of one layer is the input of the next layer. The more layers a network has, the *deeper* it gets. In the case of a feed-forward network, there are three kinds of units called input, hidden, and output units. In Figure 2.1, an example network is shown. Its most important part is the hidden layer (blue, with the bubbles representing the units in the layer), where a weighted sum of the input values is taken, followed by a non-linearity. The results are propagated to the next layer, in this case, the output layer.

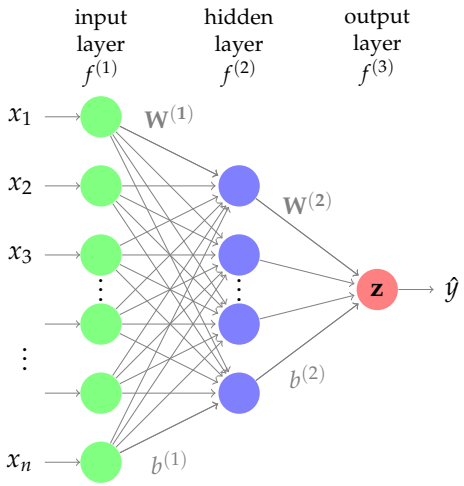


Figure 2.1: A Feed-Forward Network with three layers. It returns a single value y, for example, 1 or 0. $x_1, ..., x_n$ denotes again the input values, W are the weights $w_1, ..., w_n$ displayed as a matrix. b is the bias vector for the respective layer.

As mentioned above, each unit has a weight vector and a bias term. These are represented for the entire layer as the weight matrix $\mathbf{W}$ and the bias vector b, as shown in Figure 2.1. Each element $\mathbf{W}_{ji}$ in $\mathbf{W}$ represents the weight of the connection between the ith input unit x_i to the jth hidden unit h_j. The *hidden* layer of the feed-forward network of the examples is thus calculated as shown in Equation 2.13, using the sigmoid (σ) activation function. Note that $\mathbf{h}$ is now the *representation* of the input. The output layer then takes this representation and calculates the output.

For a binary classification task, we might have a single output node yielding the probability of one class or the other. In the case of, for example, NER, the output layer will have as many units as there are pre-defined entity types, returning a probability distribution over those types by means of a *softmax* function. The type with the highest probability is then the predicted one for the given input. The calculation for a feed-forward network with only one hidden layer might therefore look as follows, producing and estimation of the true y, often called $\hat{y}$:

$$\mathbf{h} = \sigma(\mathbf{W^{(1)}}\mathbf{x} + \mathbf{b}) \tag{2.13}$$

$$\mathbf{z} = \mathbf{W^{(2)}}\mathbf{h} \tag{2.14}$$

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z}) \tag{2.15}$$

2.3.2 Training

A feed-forward network is trained using *supervised* machine learning (Section 2.3.3), where every input example has a correct (gold-standard) output y. The goal of training is to learn the parameters $\mathbf{W^{(i)}}$ and $\mathbf{b^{(i)}}$ for each layer i to make the system's *prediction* (estimation) $\hat{y}$ as close to the true y as possible.

The distance between y and $\hat{y}$ is usually calculated using a *loss function*, e.g., cross-entropy loss. For learning, i.e., for finding the best parameters for minimizing the loss function, the gradient descent optimization algorithm is used. Since this algorithm needs to know the gradient of the loss function, i.e., a vector containing the partial derivative of the loss function with respect to each parameter, we also need an algorithm capable of calculating the gradient across all layers of the neural model. For this, in turn, the backpropagation algorithm (Rumelhart et al., 1986) is applied. We do not go into details on this and refer the reader to relevant literature, e.g., Jurafsky and Martin (2023, Chapter 5) or Goodfellow et al. (2016, Chapter 4).

2.3.3 Learning Techniques

Machine Learning can be conducted using several techniques, depending on the data. In general, there are three different approaches to learning: (i) supervised learning (ii) semi-supervised learning, and (iii) unsupervised learning.

Supervised learning assumes to have a gold standard label or class for each observation in the training and development set, i.e., for each input, there is an output. Based on this, the algorithm then aims to learn how to map from a new observation to a given correct output. The prerequisite for this setting is a (usually) human-labeled corpus available for the domain, language, and task of interest. A common supervised learning task is classification.

Semi-supervised learning can be split up into *bootstrapping* and *distant supervision* and is often applied when there are not enough data for an ML algorithm to learn from. For bootstrapping, so-called *seed patterns* (e.g., relevant entity pairs) are created, requiring very little human effort. With these patterns, documents containing them can be collected from the web or another corpus. From those and the context surrounding the given pattern, similar patterns can be deduced for collecting more documents. These, in turn, can be used to find relevant documents on the web or in other existing corpora, resulting in a new, semi-supervised dataset.

Distant supervision (Mintz et al., 2009) combines bootstrapping and supervised learning, often taking advantage of knowledge bases. It is, therefore, especially useful for REL.

Unsupervised learning works without any labeled data. It is used to detect patterns or other structures in a given dataset. Popular unsupervised methods are, for instance, clustering and dimensionality reduction. In clustering, similar data points are separated from non-similar data points, e.g., topic clustering tries to automatically find news articles belonging to the same topic within a set of articles, based on automatically extracted features. Dimensionality reduction extracts the most essential features from the underlying data by reducing the overall amount of features represented in the data. Language modeling (see Section 2.4) can also be seen as an instance of unsupervised learning: The prediction probability of a word is based on the previous word(s) and usually, models are learning these probabilities on huge corpora, where no specific labels are provided.[6]

2.3.4 Transfer Learning

Transfer learning is a method to transfer ML models to data outside their training distribution (Ruder, 2019). As shown in the taxonomy in Figure 2.2, it is used in many different contexts, not only in NLP. Usually, as described above, we assume to have data for one task and one domain and apply it within the same domain. For this, the data points are assumed to be independent of each other and identically distributed in both train and test sets (*i.i.d.* assumption).

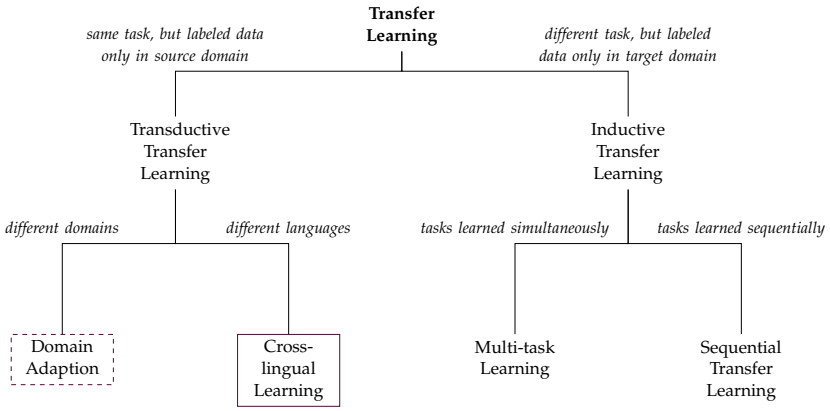


Figure 2.2: A taxonomy of transfer learning, borrowed from Ruder (2019) and slightly modified. The work described in this thesis mostly applies cross-lingual transfer learning, a sub-category of transductive transfer learning.

Thus, we can train a model on this data and expect it to perform well on unseen data from the same distribution. If we want to perform the same task in another domain, we, again, need labeled data for the respective domain. This, however, is not always possible: data annotation is tedious and costly. This is where transfer learning comes into play: It takes advantage of a related domain or task, usually called the *source* domain or task, and applies it to the *target* domain and/or task. Definition 2 shows transfer learning as defined by Pan and Yang (2010), taken over verbatim:

Definition 2 (Transfer Learning)**.** *Given a source domain D_S and learning task T_S, a target domain D_T and learning task T_T, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$.*

As shown in Figure 2.2, there are different instances of transfer learning. This thesis mostly deals with *cross- and multi-lingual learning* referring to both (i) learning cross- or multi-lingual representations of text data and (ii) transferring knowledge between a source language and a target language.

[6] In practice, the words are used as labels to the LM.

We will also encounter a change in dataset distributions between source and target data apart from language: Even though we work on user-generated data, these usually do not have the same distribution, depending on their source. This is why techniques such as *zero-shot* and *few-shot* transfer are relevant to this thesis as well.

Zero-shot Transfer describes the application of a trained model on *source* data without any modifications to *target* data. This essentially reduces the number of needed labels in the target data to zero[7].

Few-Shot Transfer assumes that there are "a few" labeled examples from the target data that can be used to train a model together with source data. Note, however, that "a few" is not formally defined, and the number of examples used in the literature can range from one to several thousand.

For a comprehensive overview of transfer learning and its different manifestations, we refer the reader to the work of Ruder (2019).

2.3.5 Models and Architectures

We continue with a brief introduction to some of the models used in the literature related to the work in this thesis. First, we will give a description of a SVM, a model from the non-deep learning era, which is used as a baseline in some of the later experiments. Next, the basics of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs) are reviewed, and finally, a short introduction to language modeling and Language Models is provided.

Support Vector Machines (SVMs) SVMs (Vapnik and Chervonenkis, 1964; Boser et al., 1992) are often applied to classification and regression tasks. A SVM constructs a hyperplane in high-dimensional space to separate classes of data points from each other. This is done by first mapping the data points into a higher dimensional space to also approach non-linearly separable data points. Second, the model learns a hyperplane that maximizes the distance to the closest training data point of any class.

SVMs need pre-defined features to learn from. These can be either hand-crafted or automatically generated, or both. Note that SVMs do not output probabilities. For a more detailed description and the mathematical background of SVMs, we refer the reader to Bishop (2006, Chapter 7) and the literature indicated in the introduction thereof.

Convolutional Neural Networks (CNNs) CNNs (LeCun et al., 1989, 1998) are most often applied to visual data like images, but have found their way into NLPs as well. Like feed-forward NNs, CNNs also contain an input layer, several hidden layers, and an output layer. In some of their hidden layers, they use a mathematical operation named *convolution* instead of the usual matrix multiplications. By sliding so-called convolutional kernels (or filters) across the input, e.g., a sequence of words, they create *feature maps*, analyzing the input step by step. These feature maps are the input to the next layer. The feature maps of the convolutional layers are commonly fed to *pooling* layers, which reduce the dimension of their input. The result can be interpreted as a "summary" of the feature maps, extracting only the most important features for the task the model is trained on. CNNs return a probability distribution over pre-defined classes using a final fully-connected layer. We refer the reader to Bishop (2006, Chapter 5.5.6) for a more detailed description of this architecture.

[7] Except when using automatic evaluation.

Recurrent Neural Networks (RNNs) RNNs (Elman, 1990), also called Elman Networks, are more commonly used in NLP tasks than CNNs because they process language (i.e., in this case text) in a sequential way, similar to humans. For this, they use *recurrent connections* to represent the context of a sequence seen before, allowing the final decision to depend not only on one word but the entire input sequence.

Simple recurrent networks have the same structure as feed-forward NNs with an input and output layer and several hidden layers. However, two differences apply. First, sequences are provided to the network step by step. Second, the recurrent step feeds outputs of a unit in a hidden layer from a preceding time step back into the same unit, acting as a kind of memory within the model. As feed-forward NNs and CNNs, RNNs are also trained using the back-propagation algorithm, but with a modification called *backpropagation through time* (Werbos, 1974; Rumelhart et al., 1986; Werbos, 1990) to account for the different time steps modeled in the NN.

Long Short-Term Memory Networks (LSTMs) LSTMs (Hochreiter and Schmidhuber, 1997) are an improved version of RNNs to account for long-distance relationships within sequences, especially texts. For this, a mechanism called *gates* proved to be useful. Gates decide which information should be kept and which can be forgotten and operate on each current input and the preceding hidden state. For more details, the reader is again referred to Jurafsky and Martin (2023), Chapter 9 on RNNs and LSTMs.

Another popular architecture in NLP is a bi-directional Long Short-Term Memory Network, a slight modification of standard LSTM. Here, a second LSTM layer is added and information can flow backward and forwards simultaneously.

Encoder-decoder networks Note that there are also *encoder-decoder* architectures, which are, for example, applied in Machine Translation (MT). These are also called sequence-to-sequence models. When using an encoder-decoder model, a sequence is fed to an encoder network, e.g., an RNN, which produces a representation of the input, i.e., an *encoding* or context vector. From this representation, the decoder generates a new sequence of arbitrary length. The decoder, too, can be any sequence model.

2.4 Embeddings and Language Models

Having discussed almost all the relevant model types for this thesis, there is now only one essential part missing: How are *words* fed to a neural model that can only "understand" numbers? And how can we model *language* in particular?

The answer to that is *word vectors* or *word embeddings*. In its very rudimentary form, a word vector is a mapping from each word w_i in the vocabulary V to a vector $\mathbf{x}_i$ of dimension $|V|$. $\mathbf{x}_i$ is 1 at exactly one dimension $\mathbf{x}_{i,j}$, all other elements of the vector are 0. For the next word $\mathbf{x}_{i+1}$, its vector is 1 at dimension $\mathbf{x}_{i+1,j+1}$, all other elements are again set to 0. Such an encoding is called *one-hot encoding*. Thus, for feeding a sequence of words to a NN, the words are first mapped to their respective vectors and then given to the network's input layer.

The above is a very naive approach and does not represent any semantics or relations between words. Therefore, better methods to create word embeddings were developed, mainly following the *Distributional Hypothesis* stating that "words that occur in the same *contexts* tend to have similar meanings"[8] (Harris, 1954; Firth, 1957), i.e., the meaning of a word is defined by its distribution in language use (Jurafsky and Martin, 2023).

[8] https://aclweb.org/aclwiki/Distributional_Hypothesis

Early work thus represented words as vectors based on matrices containing *co-occurrence counts* (Schütze, 1992, inter alia). Two widely used matrices were term-document and term-term matrices, created, for example, using Latent Semantic Analysis (LSA) (Deerwester et al., 1990). LSA was later generalized to Latent Dirichlet Allocation (LDA) (Pritchard et al., 2000; Falush et al., 2003; Blei et al., 2003).

To not only have counts of co-occurring words but more representative numbers, the tf-idf approach (Luhn, 1957; Jones, 1972) was introduced to NLP, weighing the matrix cells by their importance with respect to the documents the terms occur in. Another such weighing scheme is Positive Pointwise Mutual Information (PPMI) (Fano, 1961). It measures the probability of two events (words) occurring together compared with the probabilities of each of these events occurring independently of the other.[9]

2.4.1 word2vec & GloVe

In 2013, a more capable word representation method was introduced by Mikolov et al. (2013a,c), enabling a plethora of new applications and ideas: `word2vec`. In contrast to tf-idf or PPMI, `word2vec` vectors are short and dense: The length of the vectors does not depend on the vocabulary size anymore (often, 300 is used as the number of vector dimensions) and every value in the vector is a real-valued scalar that might also be negative.

`word2vec` is a framework comprising two algorithms to compute embeddings: `skip-gram` with negative sampling and `CBOW`. While `skip-gram` predicts if certain context words are likely to occur given the current target word, `CBOW` predicts if a word is likely to occur between two given words.

Instead of counting the frequency of neighboring words, `word2vec` trains a statistical classifier (logistic regression) to predict if a word w_j is likely to occur close to word w_i. Then, the learned weights of the classifier are taken as embeddings. For training, any text corpus of a decent size can be used: The "correct" answer (yes, w_j is likely to be close to w_i or no, it is not likely) is implicitly encoded in natural language sentences, and therefore, no manual labeling is required.

Word embeddings similar in performance, called `GloVe`, were provided by Pennington et al. (2014). `GloVe` embeddings are trained using matrix factorization and in contrast to `word2vec`, they are based on *global* co-occurrence matrices extracted from an entire corpus, not on neighboring words, which are rather *local*.

2.4.2 Language Modeling

Very similar to word embedding models, language modeling has the goal of learning the distribution of words in a corpus. This is often done by predicting the next word in a sentence, comparable to `word2vec`, i.e., modeling the probability of a word given the previous $n-1$ word(s):

$$p(w_t|w_{t-1},...,w_{t-n+1}) \tag{2.16}$$

Under the Markov Assumption[10] and using the chain rule, the probability of an entire sentence can then be approximated as follows (Ruder, 2016):

[9] Point-wise mutual information has a range between negative infinity and positive infinity. However, negative PMI values are rather unreliable if not calculated based on huge corpora. Therefore, they are replaced by zero and only positive values are taken into account.

[10] Markov assumption: $p(q_i = a|q_1...q_{i-1}) = p(q_i = a|q_{i-1})$; Applied to natural language sequences that means that for predicting the next word, only the current word matters, not the previous ones (Jurafsky and Martin, 2023).

$$p(w_1, ..., w_T) = \prod_i p(w_i | w_{i-1}, ..., w_{i-n+1}) \tag{2.17}$$

Models assigning probabilities to sequences of words are called *Language Models (LMs)*. For *n-gram based* language models, i.e., LMs that are learned from short sequences consisting of n words, the probability of a word is calculated using the corpus frequencies of the n-grams, with, for instance, Maximum Likelihood Estimation (MLE).

When using a neural network for language modeling, the input usually consists of (a representation of) a sequence of previous words, while the output units are the vocabulary. A final softmax layer is used to calculate a probability distribution over these units. The unit, that is, the word, with the highest probability is then the most likely next word.

NNs for language modeling were first introduced by Bengio et al. (2003), who proposed a simple one-layer feed-forward network for the task. They can handle a longer sequence of previous words than n-gram-based models, usually generalize better over contexts of similar words, and are also more accurate when applied in downstream tasks. However, they are not as efficient and more complex to train and run, and of course, NNs also have the disadvantage of being not *interpretable*.

Back in 2003, the computational resources for training neural LMs following Bengio et al. (2003) were not available to most researchers, with the final softmax layer being the bottleneck of the training process (Ruder, 2016). On the other hand, the models and embeddings introduced by Mikolov et al. (2013a,c) and Pennington et al. (2014) were very popular because they were both efficiently trainable and easily applicable.

Nevertheless, one problem with *static* word embeddings like those is that words with multiple meanings, e.g., "tape", are pressed into one vector, although they might describe different things depending on the context. For instance, "a tape for fixing something" is different from "a tape for recording music". Another problem is *out of vocabulary words* (OOV), i.e., words that were not seen during training the LM but occur in the test set of a task the embeddings are used for.

RNNs and LSTMs were, therefore, the next step in the development of language models, with them being able to capture certain long-range dependencies. However, one disadvantage of RNN-based models is their sequential computation which is difficult to parallelize. This stimulated further research and resulted (for now) in the invention of Transformer-based language models, which will be described next after making a short detour to cross- and multilingual language modeling.

2.4.3 Cross- and Multi-Lingual Language Modeling

Although there are more than 7,000 languages spoken all over the world, English continues to be the language most worked on in NLP (Joshi et al., 2020b; Ruder, 2022). This is due to the fact that English is spoken as first or second language in a lot of countries, but also because it was, until recently, *the* language of the internet, spoken by 14.8% of the Internet population, now apparently superseded by Chinese with 18.46% of the Internet population (Pimienta, 2022). In recent years, and particularly with the rise of deep learning, other languages are moving more towards the focus of attention since researchers are pushing towards a more inclusive NLP, not only in terms of languages but also with respect to bias, cultural background, and ethics. Nevertheless, there is still a considerable gap to bridge, especially when it comes to under-resourced languages. Even languages that are technically not low-resource, like German, French, and Japanese, do not yet reach the same level of performance for the same number of tasks, domains, and applications (Hu et al., 2020). For the remainder of this thesis, the terms *multi*-lingual and *cross*-lingual are used quite frequently:

Multi-lingual in the context of information extraction describes approaches that are applied simultaneously on several (selected) languages. For instance, a model can be both trained and tested on datasets in multiple languages.

Cross-lingual usually refers to methods that have only seen one or several languages, for example, during training (often called the *source* language(s)), and are then applied to other, non-seen languages (the *target* language(s)). The latter can be sub-categorized into languages close to the source language with respect to their language families, e.g., English and German, or very different, e.g., English and Japanese.

Multi-lingual approaches are relevant in situations where several languages (i.e., datasets in these languages) are available from the start and in a decent amount, allowing, e.g., to train only one model instead of building a system for each language separately. Countries with multiple *official* languages are, for instance, the Philippines (Filipino and English), Finland (Finnish and Swedish), or Switzerland (Italian, French, German, and Romansh).

Cross-lingual approaches, on the other side, are relevant in cases where there is not enough source data for one language. Using a similar (or even distant) source language for, e.g., training, and then applying the learned system on the desired target language might already be enough in some cases and, in the best case, drastically reduces the time and money needed to provide data and annotations. Note that these two approaches can also be mixed, e.g., when applying multi-lingual models (i.e., models trained on several languages as described in Section 2.4.5) on languages not seen during training.

2.4.4 Transformers

The main innovative idea behind Transformers, as introduced by Vaswani et al. (2017), is how the authors use the (self-) attention mechanism to improve and parallelize training. The following is loosely based on the work of Vaswani et al. (2017) and the explanations of Jurafsky and Martin (2023, Chapter 10).

The Transformer, as described in Vaswani et al. (2017), is an encoder-decoder model mapping input sequences to output sequences of the *same* length. The authors use attention mechanisms in different parts of their architecture. Additionally, they implement a concept called *positional encoding*. Both of these mechanisms help to learn relations between words in a given sequence as well as to represent time within a sequence. To describe them in more detail, we first need to understand the building blocks of Transformers.

(Self-) Attention According to Jurafsky and Martin (2023), the concept of attention was initially developed by Graves (2013) for handwriting synthesis. Graves (2013) used "soft windows" to let the network learn where it should focus for the next prediction, additionally conditioned on its previously produced output. With this, the model dynamically determines the alignment between the input text sequence and, in this case, the pen position.

Alterations of the original concept were since then used for different tasks in NLP. Bahdanau et al. (2015), for instance, applied attention in the context of machine translation. In their encoder-decoder network, they implemented a special weight matrix which learned to focus on certain parts of the input sequence when generating the (translated) output sequence, similar to the work of Graves (2013). This method is often called *self-attention*. Basically, the attention mechanism allows us to model dependencies within and between sequences, regardless of their length. Transformers are based on the idea of relying solely on (self-) attention in order to replace the common complex recurrent and convolutional operations in RNN and CNN architectures for sequence-to-sequence approaches.

Transformer Blocks A Transformer is a Neural Network with a new kind of layer whose instances are stacked on top of each other. These *Transformer blocks* (see Figure A.1), being multi-layer NNs themselves, combine four components for the encoder part: *multi-head* self-attention, feed-forward layers, normalization layers (Ba et al., 2016) and residual connections (He et al., 2016). The decoder contains the same blocks but with an additional encoder-decoder attention layer.

Residual connections are connections between layers that pass information from a lower layer to a higher one without going through any intermediate layers, accelerating learning. Normalization layers normalize the inputs such that they are kept in a certain range to make the computation of the gradient easier.

The self-attention layers allow the network to use information from any part of the context of the given sequence *up until* the current word, not only information about the current word. It also allows to compare items (words) to each other for deciding on which one to focus most for the current context.

Queries, Keys, and Values Mathematically, this comparison of items is done by using the dot product of two items, with the resulting scalar representing the similarity score of these two items. By softmax-normalizing the dot product of several item pairs with respect to the word currently under consideration, the proportional relevance of each item to the current word is calculated. As output, a sum over all inputs so far is calculated, weighted by their respective relevance. Intuitively, this represents how important the other words are for the "understanding" of the current word from the perspective of the current word. With this, the model can learn, for instance, which word or phrase in the source language is currently most relevant for the word or phrase in the target language in the case of translation. All these operations are performed independently, allowing to parallelize the computation, a big advantage of the Transformer architecture

In practice, this process is more involved since Transformers represent the relevancy of single input words in a more sophisticated way, accelerating computation by using matrix operations. Each input embedding is therefore represented as the current focus of attention (called *query* in Vaswani et al. (2017)), as the preceding input, which is compared to the current word under consideration (*key*) and also as a *value* which is used to compute the current output of the word under consideration. In the implementation of Transformers, these three perspectives are represented as learnable weight matrices Q and K of dimension d_k and matrix V of dimension d_v, which are multiplied with the input matrix X.

$$attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.18}$$

Multiple Heads Having multiple "heads", i.e., parallel self-attention layers, allows the model to focus on different aspects of a sequence at the same time, not having to choose one only.

Positional Encodings In natural languages, word order is important. However, in contrast to, e.g., RNNs, Transformers look at the input sequence without any information about the words' positions in the sequence – they do not "see" the sequence word by word, but all at once. Vaswani et al. (2017), therefore, add another kind of embedding, called positional embeddings to encode the position of the words with respect to the other words in the given sequence.

The Transformer architecture is now used in many different fields, e.g., in computer vision, audio, and multi-modal processing. However, in the next section, we will focus on its applications in NLP, more specifically, Transformers as LMs.

2.4.5 Transformer-based Models

This section describes language modeling using Transformer-based models, with a focus on the models used and referred to in this thesis. With the era of Transformer-based LMs, the concepts of *pre-training* and *fine-tuning* entered the lingo of researchers and practitioners.

Pre-Training is often utilized in the same sense as *training from scratch*, meaning that a freshly initialized architecture without any knowledge is newly trained on some data, learning new weights [11]. In the case of LMs, pre-training is often done using a *language modeling objective*, e.g., as we have seen, next-word prediction. This procedure usually needs a large amount of data to be successful and several days, weeks, or even months to train, depending on the availability of computing resources.

Fine-Tuning , in contrast, assumes an already trained model, which is then further tuned or refined on a downstream task. The fine-tuning process can also include a repetition of the pre-training task or other learning techniques, like self-training. Specific layers of a model, e.g., the embedding layer, can be frozen so they are not changed again during fine-tuning.

Language Modeling with Transformers For language modeling, Transformers can be used similarly to, for instance, RNNs. On a text corpus of decent size, the model is trained auto-regressively to predict the next word in a sequence. In contrast to other methods, the Transformer has access to much more information. For example, the model has information about the correct previous part of the sequence as well as knowledge about positions and about every word up until the one currently under consideration. Thanks to the attention mechanism, it also knows which words to focus on to generate the next word. Transformers are still limited in that they only see the *previous* context and not all of the given sequence. This does not matter so much in, e.g., summarization or machine translation (Jurafsky and Martin, 2023). However, for more "fine-grained" tasks in NLP, like sequence labeling (for instance, NER) or sequence classification, it can be useful to take a look at the rest of the context as well.

`BERT` *Bi-directional encoders* improve upon this shortcoming by letting the self-attention mechanism attend to all words in the given input sequence, not only the ones up until the current word. The first work introducing this idea was published by Devlin et al. (2019). This model produces *contextualized* representations of words, i.e., depending on its context in a sentence, the same word can have different representations (embeddings). The main components of the Transformer architecture underlying `BERT` are the same as before. This time, however, the part of the sequence *after* the current word is also considered, allowing the model to access information from two directions. `BERT` further uses WordPiece tokenization (Schuster and Nakajima, 2012; Wu et al., 2016), exactly as Vaswani et al. (2017), resulting in a sub-word vocabulary of 30,000 subword-tokens[12] (Devlin et al., 2019).

Since `BERT` can see both the left and the right context of the current word, the next-word-prediction task has become trivial. Therefore, Devlin et al. (2019) use a different training strategy. They train `BERT` on two unsupervised tasks, (i) Masked Language Modeling (MLM), and (ii) Next Sentence Prediction (NSP). For MLM, as the name suggests, 15% of the input

[11] Of course, models can be trained directly on downstream tasks from scratch as well.

[12] Subword-tokens are tokens that are not necessarily "whole" words. Only frequent words are kept in their original version, and rare words are split up into more frequent subwords, allowing the creation of a vocabulary of reasonable size and, at the same time, getting rid of OOV words since these can be constructed from sub-words.

subword-tokens are *masked*, and the model has to predict the actual word for the masked position, similar to a fill-in-the-blank task[13]. For predicting the actual token, the authors use again the cross-entropy loss.

For the second objective, NSP, Devlin et al. (2019) extract consecutive sentences from a corpus and train the model on a binary task for predicting if the second sentence is the "next" sentence or not. Both tasks are learned at the same time.

Devlin et al. (2019) first *pre-train* their model on the described language modeling objectives and then *fine-tune* it on several down-stream tasks, for instance, natural language understanding using the GLUE dataset (Wang et al., 2018a). An overview of the pre-training and fine-tuning procedure of `BERT` can be seen in Figure A.2, as well as some more details on `BERT`.

Note that LMs also introduce potential harms. These are briefly described in Appendix A.2.

[13]This is also often called a Cloze task (Taylor, 1953).

Chapter 3

Related Work

This chapter reviews the advances in multi- and cross-lingual Information Extraction (IE) as well as the current state-of-the-art in biomedical IE. Then, these two fields are combined and the current state of cross- & multi-lingual IE in biomedical NLP is revisited. All works are reviewed focusing mainly on document classification and entity extraction since these tasks are relevant for the rest of this thesis. Finally, since data availability is an essential topic in this thesis, the currently existing datasets will also be described. For completeness, standard tools and databases in biomedical NLP are added in Appendix A.4.

3.1 General Cross- & Multi-Lingual Information Extraction

Cross- and multi-lingual IE has been worked on for quite some time. A prominent player in advancing research in multi-lingual IE is and was the series of "Cross-Language Evaluation Forum" (CLEF) workshops and shared tasks[1]. The organizers set the real beginning of research into multi- and cross-lingual language processing to the 1990s (CLEF, 2016). Already in the 1960s, cross-lingual information retrieval was a topic in library sciences. The rise of the Internet then accelerated interest and work in multi-lingual IE to improve access to relevant information from multiple languages (CLEF, 2016). The first widely known workshop on multi-lingual IE outside of Europe was organized in 1999 by the National Institute for Informatics (NII) in Japan, focusing on Asian languages, mostly Japanese, Chinese, and Korean. The workshop series was called "NII Testbeds and Community for Information Access Research", in short, NTCIR, and is still active today, holding its 17th version in December 2023 at NII[2].

Back then, multi-lingual research was mostly supported by four types of resources: dictionaries, parallel corpora, comparable corpora, and machine translation programs (CLEF, 2016). This has changed in the last years with the introduction of (monolingual) vector representations of words (Mikolov et al., 2013a,b,c; Pennington et al., 2014), multi-lingual word representations (Al-Rfou et al., 2013; Grave et al., 2018) and the rise of contextualized language models (Vaswani et al., 2017; Peters et al., 2018; Devlin et al., 2019), which soon were available for multiple languages, too (Devlin et al., 2019; Conneau et al., 2020).

3.1.1 Static Embeddings

Embeddings, contextualized or not, are the *de facto* basis for most tasks in NLP today, and often the focus of the overall processing pipeline. Before large language models became feasible, various ways of creating embeddings for several languages were proposed, which will be briefly discussed below.

Many embedding generation methods rely on parallel or comparable data. Parallel data refers to direct translations between source and target languages, while comparable corpora

[1] http://www.clef-initiative.eu/; CLEF was renamed in 2010 into "Conference and Labs of the Evaluation Forum".

[2] https://research.nii.ac.jp/ntcir/ntcir-17/index.html

consist of data in different languages that are similar but not exactly the same. An example of the latter are Wikipedia articles in German and English about the same topic. Often, parallel corpora (Yu et al., 2018) or dictionaries (Mayhew et al., 2017; Xie et al., 2018) are therefore used for mapping several languages into a common space by creating fixed-dimensional sentence embeddings and penalizing those that are too far from each other during a translation task.

Another line of work in this context are approaches that take mono-lingual embeddings and map them into a common multi-lingual space. This can be, again, achieved using several methods, e.g., mapping the source language space to the target language space, or mapping both spaces into a common one, always maximizing the similarity between the single word vectors. Similarly, other approaches create multi-lingual *sentence* embeddings by mapping (sometimes already existing) mono-lingual embeddings, for instance trained with the `GloVe` or `word2vec` algorithms, from different languages into one common embedding space (Schwenk and Douze, 2017) or by using either `CBOW` or a bi-directional Long Short-Term Memory Networks (bi-LSTMs) for encoding the sentences (Conneau et al., 2018).

Further strategies are based on adversarial learning (Keung et al., 2019), phonological representations of characters (Bharadwaj et al., 2016), semantic role labeling Akbik et al. (2016) or universal schemata (Riedel et al., 2013) to create language-agnostic language embeddings. Lin et al. (2017) experiment with aligning Wikipedia articles in different languages for better results in multi-lingual REL. A mapping of bilingual word embeddings from target to source language using a learned linear mapping as introduced by Mikolov et al. (2013b) is applied by Ni and Florian (2019). Various other approaches make use of bilingual dictionaries (Ni and Florian, 2019) or graphs (Kim et al., 2014), extract relation embeddings (Lin et al., 2017) or extract independent sentence embeddings per language (Wang et al., 2018b).

3.1.2 Neural Models

In terms of models, researchers experimented with different methods as well. Note that these do not differ much compared to mono-lingual approaches. Schwenk and Li (2018), for instance, apply both a Multi-Layer Perceptron (MLP) and a CNN for document classification. Conneau et al. (2018) use a basic one-layer feed-forward neural network on top of their sentence embeddings. Keung et al. (2019) as well as Dong and de Melo (2019) and Hu et al. (2020) explore `mBERT` (Devlin et al., 2019). `mBERT` is also used by Eronen et al. (2023), as well as `XLM-RoBERTa`. Hu et al. (2020) further experiment with `XLM` and `XLM-RoBERTa` (large).

For NER, frequently used models and architectures are, amongst others, Conditional Random Fields (CRFs) (Lafferty et al., 2001), employed, for instance, in work by Ni et al. (2017), or Maximum Entropy Markov models (MEMM) (McCallum et al., 2000), also explored by Ni et al. (2017). Equally popular are bi-LSTMs, often in combination with CRFs (Pan et al., 2017; Xie et al., 2018; Bari et al., 2020) or equipped with an additional linear classifier (Adelani et al., 2021), or combinations of CRFs, bi-LSTMs and CNNs (Ma and Hovy, 2016; Rijhwani et al., 2020; Adelani et al., 2021). For the more traditional approaches and also in combination with more recent architectures, gazetteers are a popular feature (Rijhwani et al., 2020; Adelani et al., 2021). Recently, mostly `BERT` or `mBERT` (Lauscher et al., 2020; Adelani et al., 2021, inter alia) or `XLM-RoBERTa` (Lauscher et al., 2020; Ebrahimi and Kann, 2021; Adelani et al., 2021; Kulkarni et al., 2023; Tan et al., 2023) are used, in one case `BERT` initialized with `mBERT` embeddings and then re-trained from scratch (Arkhipov et al., 2019). The Transformer-based models are further often topped with a CRF or with a bi-LSTM plus a CRF (Tedeschi et al., 2021).

Especially `XLM-RoBERTa` is a popular choice for cross- and multi-lingual tasks. The model is based on `BERT` but with modifications already established in `XLM` and `RoBERTa` models.

`XLM` The model introduced by Conneau and Lample (2019) slightly modifies the original MLM objective of `BERT`. Further, for tokenization, it uses a shared multi-lingual vocabulary

created with Byte Pair Encoding (BPE) (Sennrich et al., 2016). Conneau and Lample (2019) further introduce "translation language modeling" (TLM) for cross-lingual training and combine it with their adapted version of MLM. Using parallel corpora in different languages, they concatenate parallel sentences as input and randomly mask words in the source and the target sequence. With this, they are able to outperform `mBERT` on several multi-lingual tasks.

`RoBERTa` This is a mono-lingual model for English proposed by Liu et al. (2019) and is, again, based on `BERT`. The authors remove the NSP pre-training objective and further extend the pre-training corpus of `BERT` with more English corpora. With this and longer pre-training than Devlin et al. (2019), the authors arrive at a more robust version of `BERT`, for both the `base` and `large` version.

`XLM-RoBERTa` Conneau et al. (2020) introduced a combination of both `RoBERTa` and `XLM`, with the same implementation as `RoBERTa`. It is trained on data in 100 languages, but, in contrast to `XLM`, it does not use the TLM technique, but rather original MLM on sentences from the same language. The authors use the SentencePiece tokenization algorithm[3] for language-agnostic tokenization. `XLM-RoBERTa` architecture contains 24 Transformer layers, with a hidden layer size of 1,024.[4] It is trained on 2.5 TB of a cleaned multi-lingual version of CommonCrawl, following the work of Wenzek et al. (2020). `XLM-RoBERTa` outperforms the above-mentioned models and shows comparably strong performance on low-resource languages.

3.1.3 Approaches

The tasks tackled in this thesis are approached in various ways throughout the literature. Some of the most prominent ones are described in the following.

Zero-Shot One of the most often employed techniques for cross-lingual IE is zero-shot transfer. This is due to the fact that there are usually very few resources, mostly not enough to train on, but only for evaluation. For example, Schwenk and Li (2018) apply the zero-shot method by using multi-lingual word embeddings provided by Ammar et al. (2016) to address document classification.

Few-Shot Some authors tackle the limitations of "simple" zero-shot transfer and advocate for more research into few-shot transfer, particularly for under-resourced languages (Lauscher et al., 2020). Others experiment with various amounts of added target language data (Hennig et al., 2023) to improve performance and gain insights into the amount of (annotated) target data needed to achieve a decent performance.

Annotation Projection Another line of work relies on translations (Mishra and Haghighi, 2021) or token alignments to transfer, e.g., entity annotations between languages (Ni et al., 2017; Xie et al., 2018). Kim et al. (2014), for instance, work with parallel corpora in English and Korean to project annotations from source to target language. Translations can also be used in other ways, e.g., by taking translations of the training data, or by taking translations of the test data (and the monolingual embeddings of those) as done by Conneau et al. (2018, inter alia).

[3] https://github.com/google/sentencepiece

[4] The authors do not mention the number of attention heads, but we assume it to be 16, like in $\texttt{BERT}_{large}$.

Self-learning Dong and de Melo (2019) apply self-learning in multi-lingual text classification by incorporating `mBERT`'s predictions on non-English data into training on English data. This method is similar to an approach proposed by Eisenschlos et al. (2019) and also evaluated by Hu et al. (2020), who label a "small" set of 1,000 examples of the target language and feed it to an LM[5]. Other publications report experiments incorporating specific sample selection strategies (Ni et al., 2017; Pan et al., 2017) to the self-learning scheme, where a model performs a silver labeling of target data and gets high-confidence samples fed back during the next training iteration. Various works also explore the selection of specific samples to learn from (Prelevikj and Zitnik, 2021) and integrate self-attention (Xie et al., 2018) or attention over character sequences (Bharadwaj et al., 2016) in their models.

Adversarial Training Adversarial training is usually done using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), where a *generator* produces "fake" examples, while a *discriminator* has to distinguish these examples from the real ones. Dong et al. (2020) use this approach for document classification. In contrast to the work done by Keung et al. (2019), this time, the discriminator has to learn to distinguish between original and perturbed examples, where perturbing means that the embeddings of the non-English word from the target language replace the embedding of an English word in the document. Adversarial feature adaption is used in works of Zou et al. (2018) and Wang et al. (2018b), where the authors use a GAN to generate language-agnostic features that can be transferred between two languages.

Translations Another promising line of work is translation of source or target data (Faruqui and Kumar, 2015; Nag et al., 2021; Hennig et al., 2023) and back-translation (Faruqui and Kumar, 2015) or translation of prompts (Chen et al., 2022a), as well as using translations based on similarity (Artetxe and Schwenk, 2019). Kolluru et al. (2022) aim to improve translations and annotation projections by biasing a sequence-to-sequence model to translate sentences similar to a reference translation.

Continued or improved pre-training Continued pre-training (Ebrahimi and Kann, 2021; Fan et al., 2021) or improved pre-training, for instance, by adding more or different pre-training objectives is another way to improve performance in multi-lingual NER. For instance, Mishra and Haghighi (2021) add a translation pair prediction objective to the standard `mBERT` objective, which helps the model to learn if a sequence is a valid translation of a source sequence or not.

Adapting model specifics Some works augment Transformer-based models, e.g., by extending their existing vocabulary (Wang et al., 2020; Ebrahimi and Kann, 2021; Adelani et al., 2021) or by creating a new, more language group-specific vocabulary via language clusters (Arkhipov et al., 2019; Chung et al., 2020) to balance the trade-off between sharing sub-tokens across languages and language-specific tokens.

Apart from different architectures, training strategies, or combinations of models, other work considers adapting the original attention mechanism (Vaswani et al., 2017). For instance, both Lin et al. (2017) and Wang et al. (2018b) apply cross-lingual attention between relation embeddings, sometimes also in combination with language-specific attention (Wang et al., 2018b).

Adapters A method employed quite often recently is language adaptive fine-tuning, e.g., in the form of adapters (Pfeiffer et al., 2020), often used for NER (Alabi et al., 2020; Adelani et al., 2021; Ebrahimi and Kann, 2021; Kulkarni et al., 2023, inter alia).

[5]This can also be seen as an instance of few-shot learning.

Multi-Task Learning Another approach gaining more attention in recent years is multi-task learning (Caruana, 1993) to exploit predictions of one task to help disambiguation in other tasks. For instance, Sanh et al. (2019b) combine NER, entity mention detection, coreference resolution, and REL in one model.

Prompting A relatively new idea is pursued by Chen et al. (2022a) who try to extract relations using prompting in various settings and in different languages.

3.1.4 Summary

Summarized we find that most approaches are carried out in a low-resource setting, since usually, the target data size is too small to be trained on. Because of this, zero- and few-shot approaches are common, often in the form of self-learning, where noisily labeled target data gets fed back into the model.

Before the introduction of multi-lingual encoders like `mBERT` or `XLM-RoBERTa`, cross-lingual transfer was mostly improved by aligning or projecting annotations in different languages, or by translation, either during training or during testing. These methods might be "older" but are still relevant. In general, and as also shown by Yarmohammadi et al. (2021), Chen et al. (2022b) and Eronen et al. (2023) (amongst others), there is no solution that answers all cross- or multi-lingual questions in IE. Each task, language, and data setting calls for a different setup. Therefore, available data, encoders, decoders, external knowledge, and domains need to be carefully investigated and geared to each other to achieve a good performance. Furthermore, as Yarmohammadi et al. (2021) emphasize in their work, it is important to explore combinations of different methods, introduced both before and after multi-lingual encoders, to get the best possible results.

3.2 Information Extraction in Biomedical NLP

This section describes (mono-lingual) IE in biomedical and clinical NLP to demonstrate the challenges researchers are currently facing in this domain and how they try to solve them. The focus of this section is on Adverse Drug Reactions (ADRs) to set the work presented in this thesis into context.

Information extraction and, in general, clinical and biomedical text mining methods have been used for quite some time to distill knowledge from documents related to healthcare. The methods are usually borrowed from the general domain and adapted to specific vocabulary and smaller resources. Biomedical and clinical datasets show, similar to general domain data, a wide variety in style, length, jargon, and other characteristics. There a medical records, e.g., Electronic Health Records (EHRs), written differently in every hospital and laboratory, scientific publications about the newest insights into a variety of topics, public health or treatment guidelines, and finally, social media, e.g., patient fora or Facebook user groups. In each of these areas, medical information is verbalized in different ways and with different goals (Rodriguez-Esteban, 2009). See Figure A.3 for an overview of common data sources.

Usually, document classification strategies are used to categorize abstracts, user posts or paragraphs into specific topics, for instance, health hazards depending on region (Collier et al., 2008), reports about adverse drug reactions (Alimova et al., 2017), ICD codes (Silvestri et al., 2020; Ibrahim et al., 2021), or if someones is talking about anxiety issues online (Shen and Rudzicz, 2017). Named entity recognition or span categorization in biomedical texts is geared towards the detection and extraction of specific terms, e.g., proteins, cell types, medication names, devices, chemical compounds, diseases, symptoms, etc. (Rodriguez-Esteban, 2009), but also social determinants of health (Lybarger et al., 2023) can bring insights on and for patients.

The relations between these terms are of interest as well. For example, researchers extract drug-drug (Thomas et al., 2011; Bobic et al., 2012) or protein-protein interactions (Tikk et al., 2010; Bobic et al., 2012), but also relationships between drug and adverse reactions (Gurulingappa et al., 2012b), just to name a few.

Most work described below concerns itself with texts in English, due to the vast majority of works being on English data, but whenever possible, some pointers to works on non-English datasets with respect to ADRs are given as well. Multi-lingual approaches will then be discussed in Section 3.3. See Appendix A.4 for more details on commonly used biomedical databases and ontologies.

3.2.1 Data

Leaman et al. (2010) were one of the first to apply biomedical NLPs methods to social media data, in this case, the patient forum DailyStrength[6]. One year later, Chee et al. (2011) also worked on the classification of forum posts from Health & Wellness Yahoo! Groups. Data from DailyStrength was, for example, (re-) used by Patki et al. (2014), Nikfarjam and Gonzalez (2011), Sarker and Gonzalez (2015) and many others. Similarly, Metke-Jimenez et al. (2014) and Karimi et al. (2015) work on data from the English patient forum AskAPatient.

Different from that, Harpaz et al. (2010) work with voluntary ADR reports provided by the U.S. Food and Drug Administration. Similar work was done on EHRs, e.g., by Friedman (2009) and Aramaki et al. (2010), on MEDLINE abstracts (i.e., medical case reports) (Gurulingappa et al., 2012b; Huynh et al., 2016) and using PubMed articles (Shetty and Dalal, 2011). The dataset created by Gurulingappa et al. (2012b) became a widely used benchmark corpus, especially for relation extraction, see, e.g., the work of Arannil et al. (2023) and others.

Note that the EHRs in the work of Aramaki et al. (2010) were written in Japanese and, therefore, this work presents one of the first (published) works on non-English data. Ginn et al. (2014) work with Twitter messages, where they annotate tweets with drug and reaction mentions and map them to Unified Medical Language System (UMLS) concepts. They are one of the few who also published their dataset. Sarker and Gonzalez (2015) also mine Twitter for messages containing ADRs. From these data, they provided a Twitter corpus in the first SMM4H task in 2016 (Sarker et al., 2016), which was then used by Huynh et al. (2016) and others. With the work of Sarker et al. (2016), the series of the Social Media Mining for Health (SMM4H) shared tasks was started and has been held since then every year, reflecting the current state-of-the-art in biomedical text processing on user-generated texts, often with a task on ADRs as well. Alvaro et al. (2017) provide a new corpus called TwiMed which combines English Twitter messages with PubMed. In 2017, another non-English social media dataset was published: Alimova et al. (2017) collected Russian drug reviews and labeled sentences with the classes *Indication, Beneficial effect, Adverse drug reaction, Other* (see Section 3.4). Combi et al. (2018) work on Italian reports on ADRs, but unfortunately never released the data. Thompson et al. (2018) released a new English corpus called PHAEDRA, containing MEDLINE abstracts annotated with ADRs, drug mentions, and drug interactions, which are further linked to concepts in appropriate databases, such as Medical Subject Headings (MeSH) and SNOMED-CT (see Appendix A.4).

In the fourth SMM4H shared task in 2019, Weissenbacher et al. (2019) provided the participants with tweets for classification, span detection, and linking with respect to ADRs. The tweets, however, were not newly collected but re-used from previous shared tasks (Sarker et al., 2018). In contrast, the fifth version of SMM4H was the first one that provided also non-English tweets for classification, namely in French and Russian (the datasets are described in more detail in Section 3.4).

[6] https://www.dailystrength.org/

Lee et al. (2020) release the BioBERT model and run it on various datasets, amongst others on the English EU-ADR corpus (van Mulligen et al., 2012), which contains MEDLINE abstracts annotated with entities and relations with respect to ADRs. Finally, Scaboro et al. (2022) provide an English benchmark dataset called SNAX which is intended to serve as a test bed for systems with respect to speculations and negation in the detection of ADRs. Note that the datasets used in this thesis are described separately in Section 3.4.

3.2.2 Methods & Models

The algorithms and models very much follow the development in general IE, but are mostly adapted to the domain. As mentioned, databases and ontologies play an important role in biomedical text processing, and therefore, they are also extensively used in the methods described below.

String Matching The very first approaches to the detection of ADRs were based on string matching and domain-specific lexicons (Leaman et al., 2010; Chee et al., 2009a,b; Friedman, 2009; Metke-Jimenez et al., 2014). Often, these lexicons were extracted from databases like Medical Dictionary for Regulatory Activities (MedDRA) or SIDER (Kuhn et al., 2016), but a common approach also relied on sentiment lexicons since ADRs are usually described with a more negative sentiment. Jiang and Zheng (2013) use a software called MetaMap (Aronson, 2001; Aronson and Lang, 2010) to detect effects of medication intake. Other approaches were based on patterns and/or rules (Nikfarjam and Gonzalez, 2011), while Harpaz et al. (2010) aimed to find drug-drug reactions using the Apriori algorithm (Agrawal, 1994), a method to detect associations in large databases.

Machine Learning Approaches After being used in general NLP tasks for a while, SVMs and (Multinomial) Naive Bayes approaches became popular in biomedical NLP, too (Chee et al., 2011; Gurulingappa et al., 2012b; Yang et al., 2013; Jiang and Zheng, 2013; Patki et al., 2014; Ginn et al., 2014). These were also based on specialized lexicons and hand-crafted features such as word frequencies (Chee et al., 2011), Part-of-Speech (POS) tags (Gurulingappa et al., 2012b) or n-grams (Patki et al., 2014). Cami et al. (2011) trained a logistic regression model on database entries from 2005 to predict new drug-ADR associations on data from 2010. Gurulingappa et al. (2012b) further showed that a Maximum Entropy classifier worked well for binary classification of MEDLINE documents, while the same was shown for Twitter data by Jiang and Zheng (2013).

Going further than classification, Aramaki et al. (2010) used CRFs to identify symptoms and drugs in Japanese EHRs and a SVM as well as a pattern-based approach to classify the relations between drugs and symptoms. Segura-Bedmar et al. (2014) used the tool TextAnalytics and created gazetteers to extract both drug mentions and ADRs. For the latter, they used MedDRA and CIMA[7]. Instead of CIMA, Metke-Jimenez et al. (2014) used UMLS to find documents containing relevant terms.

Sarker and Gonzalez (2015) experimented with data from DailyStrength combined with tweets and the ADRCorpus provided by Gurulingappa et al. (2012a). Using again SVM, Maximum Entropy, and Naive Bayes classifiers for binary classification, they experimented with three datasets, two of which are sources from social media, while the third one is based on MEDLINE.

Nikfarjam et al. (2015a) also applied CRFs on previously successful features but now in combination with word embeddings trained with word2vec. Other authors used ensembles of decision trees, e.g., Rastegar-Mojarad et al. (2016). For the Russian drug review dataset,

[7]CIMA is a Spanish online medication information platform. http://www.aemps.gob.es/cima/

Alimova et al. (2017) used SVMs based on hand-crafted features and `word2vec` embeddings for classification and added class weights to account for the class imbalance.

Combi et al. (2018) aimed to link descriptions of ADRs to their respective MedDRA terms and apply the tool MagiCoder, which is a system that scans the texts word by word and tries to detect tokens that belong to known MedDRA entries while performing various pre-processing tasks such as stemming. The authors of PHAEDRA (Thompson et al., 2018) trained NERSuite[8] whose main NER component is a CRF[9]. Next to SVMs for sentence classification, Zolnoori et al. (2019) used the clinical Text Analysis and Knowledge Extraction System (cTAKES, Savova et al. (2010)) for NER, which is again, a dictionary lookup algorithm.

Deep Learning Approaches Huynh et al. (2016) proposed three neural models for the classification of ADR-related documents: A convolutional RNN, a CNN with attention, and a recurrent CNN. The experiments are conducted on tweets (Sarker et al., 2016) and MEDLINE (Gurulingappa et al., 2012b). Their methods improved upon previous ML approaches but performed worse on the social media data than on the MEDLINE corpus, while a simple CNN outperformed the more complex versions. Similar work was done by Lee and Uzuner (2020) with a standard RNN for concept normalization on the TAC2017 shared task dataset, the FDA shared task dataset, and the SMM4H 2019 data, all English.

The participants of the SMM4H shared task 2018 mostly used CNNs or RNNs in combination with word embeddings (Weissenbacher et al., 2018, 2019). Uzuner et al. (2020) summarized the participating systems in the n2c2 shared task from 2018 and reported similar trends as shown in other related work: RNNs were a popular choice among participating systems, while also combinations of traditional and deep ML models were successful. They further stated that ensembling and post-processing helped in increasing the final performances on the test set.

Yang et al. (2020) incorporated knowledge from SIDER into the embeddings of their architecture for NER, a recurrent CNN, and found that precision improved but recall decreased. Other combinations of, for example, LSTMs and CRFs were used, e.g., by Sutphin et al. (2020).

Transformer-based Approaches With the rise of `BERT` and companions, domain-specific models were published as well. Therefore, some often used biomedical models are described briefly to highlight the differences between training methods and underlying datasets. Note that all of these are trained on English data and based on the above described `BERT` architecture. To the best of our knowledge, no multi-lingual biomedical model exists so far.

`BioBERT` (Lee et al., 2020) is a `BERT` model pre-trained as described above. The model pre-training is then continued[10] using abstracts and full texts from PubMed (4.5 billion words), a database for scientific papers[11].

`BioClinicalBERT` (Alsentzer et al., 2019) was initialized with `BioBERT`[12] and fine-tuned on all available data in the MIMIC-III corpus (Johnson et al., 2016), which contains EHRs from ICU patients.

`BioRedditBERT` (Basaldella et al., 2020) was initialized with `BioBERT` and continuously pre-trained on the COMETA corpus (Basaldella et al., 2020), a text collection containing Reddit posts related to medical topics in colloquial language with approximately 300 million tokens.

[8] `http://nersuite.nlplab.org/`

[9] CRFSuite (Okazaki, 2007), `http://www.chokkan.org/software/crfsuite/`

[10] This is called *continual pre-training* and uses the pre-training objectives, in contrast to fine-tuning, which uses task-specific objectives for learning.

[11] `https://pubmed.ncbi.nlm.nih.gov/`

[12] Both the `BioBERT` paper and model were already published online before being accepted for *Bioinformatics*.

PubMedBERT (Gu et al., 2021) is a model trained from scratch on 3.1 billion words from filtered PubMed abstracts, following the assumption that pre-training on domain-specific data is better in terms of performance than continual pre-training or fine-tuning (Gu et al., 2021).

The 2019 edition of the SMM4H shared task was characterized by an adoption of BERT-based models for the classification, extraction, and linking of ADRs (Weissenbacher et al., 2019). However, the organizers also noted that even with Transformer-based models, the tasks are still challenging: The best system achieved an F_1 score of 0.646 only for the classification of ADR documents.

With the SMM4H 2020 version, also non-English BERT models came into play. Both Miftahutdinov et al. (2020) and Gusev et al. (2020) used ensembles of Russian BERT models and additional training data to create the winning systems for the Russian tweets. For the French data presented at SMM4H 2020, SBERT (Reimers and Gurevych, 2019) and DistilBERT (Sanh et al., 2019a) in combination with class weights won the challenge (Gencoglu, 2020). Note, however, that a logistic regression approach using hand-crafted features (Tanguy et al., 2020) came very close to the results on the difficult French tweets[13], where the best team achieved an F_1 score of 0.17.

(Lee et al., 2020) test their BioBERT model on EU-ADR, where they show a slight improvement over standard BERT with their model trained on PubMed abstracts and PubMed Central full articles (F_1 score for relation extraction is 86.51). Tutubalina et al. (2020) provide a Russian drug review corpus (details in Section 3.4) and pre-trained models derived from mBERT. They pre-train a BERT-based model, which they call RuDR-BERT, on unlabeled data and apply it for classification and NER.

Haq et al. (2021) provided new results on the ADE corpus of Gurulingappa et al. (2012b) by using BioBERT. However, in a study conducted by Portelli et al. (2021), the authors demonstrated that the best models for the detection of ADRs mentions are (as of 2021) PubMedBERT (Gu et al., 2021) and SpanBERT (Joshi et al., 2020a), showing that span-based pre-training has a big effect on precise detection of ADR mentions and also that in-domain pre-training, even if it is not on social media texts, can still outperform general domain models.

Raval et al. (2021) introduced a new technique for the classification and extraction of ADRs by using a generative model, namely T5 (Raffel et al., 2020). They experimented with several different datasets created from social media (SMM4H 2018-en, SMM4H 2020-fr (Twitter), CADEC, ADE-v2, WEB-RADR) and trained the model on all tasks (depending on the dataset) at the same time. As baselines, they tried several BERT variants (BioBERT (Lee et al., 2020), BioClinicalBERT (Alsentzer et al., 2019), SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021), SpanBERT (Joshi et al., 2020a)), but also combinations of BERT models with a final CRF layer. To account for the different dataset sizes, they introduced *proportional mixing*, i.e., sampling in proportion to the size of the corpus. The authors applied this sampling scheme together with *temperature scaling* as shown by Raffel et al. (2020) and others to balance the task and dataset. With this, they improved upon other work in both classification and extraction.

DeepADEMiner is a complete pipeline for extracting and normalizing ADRs in tweets published by Magge et al. (2021). It includes RoBERTa for classification, an NER component for span extraction, and either a fastText or BERT classifier for normalization. The authors showed, with the best performance of the complete pipeline resulting in an F_1 score of 0.34, that there is a lot of room for improvement even on English Twitter data, which is one of the domains most worked on.

Zhu and Jiang (2021) investigated semi-supervised techniques for ADR classification, mixing labeled and unlabeled tweets for training a BERT-based model trained from scratch on tweets. They first generated a silver-annotated training set by applying a classifier trained on

[13]The data contained only 1.6% positive tweets in both train (2,426 tweets overall) and test set (607 tweets overall).

labeled data to unseen data. Further, they added what they call "consistency regularization" to make certain that the model stays consistent in its predictions even when adding new data. They then showed that their techniques outperform a baseline relying only on labeled data, i.e., pseudo-labeled data seems to be better than gold-labeled samples, of which there are only a few.

Huguet Cabot and Navigli (2021) provided a new LM called `REBEL` which is based on the sequence-to-sequence model `BART`. They framed relation extraction similar to a translation task by feeding a sequence to a model, which then returns a triplet of entities and a relation, i.e., a translation of text into triplets.

Similarly, Paolini et al. (2021) proposed a method that is based on the idea of translation as well. On datasets such as the ADE dataset (Gurulingappa et al., 2012b), they demonstrated joint entity and relation extraction by encoding the given input sequence with the relevant information with respect to entities and relations, which are then decoded into the desired structured output using the `T5` model.

Scaboro et al. (2021) and Scaboro et al. (2022) investigated the effect of negation in the context of ADRs and showed that adding negated samples and a specific negation detection component can help in improving the robustness of `BERT`-based models for both the classification and extraction of ADRs.

A question-answer-based approach for the detection of ADRs was proposed by Arannil et al. (2023), working in a similar way as the approach of Raval et al. (2021). The authors used the generative model `T5` (Raffel et al., 2020) in combination with a Question & Answering (QA) task in a multi-task scenario to extract entities and relations separately. In a second approach, they fed questions and context to a model, which then returns the extracted ADRs and drug mentions in one go. However, they found the separate approach to be more successful.

3.2.3 Challenges

Although the above does not show every technique used for approaching the task of classification, detection, and relation extraction with respect to ADRs, it still demonstrates the development of the methods over time. Some challenges, like reasoning with context, showed improvement with the introduction of Transformer-based models, but various challenges remain.

Label Imbalance One of the challenges faced often in biomedical NLP is the imbalance of data. This was, for example, reported in the work of Ginn et al. (2014), who collected tweets based on drug keywords. The data was balanced by drug mention, but not in terms of ADRs, i.e., only 11% actually contained ADRs. Similarly, in the Japanese EHRs used by Aramaki et al. (2010), not even 8% contained ADRs. To counteract the imbalance, Ginn et al. (2014) test their models in different configurations, showing that the more imbalanced the data, the worse the performance of the models, in this case SVM and Naive Bayes.

Sarker and Gonzalez (2015) combine datasets from different sources to expand the number of (positive) training examples. The models trained on data from Twitter and DailyStrength benefit from the combination, but when adding social media data to the MEDLINE dataset, the performance does not change significantly.

Magge et al. (2021) also emphasize the "natural" label imbalance observed in the data as a major obstacle to achieving good performance, even when experimenting with undersampling techniques and lowered thresholds for classification.

Data Collection & Annotation Methods Another issue with respect to generalization is the collection method. Especially for social media data, researchers usually use keywords to harvest Twitter or other social networks as described by Ginn et al. (2014) or Metke-Jimenez et al.

(2014) who both rely on a restricted set of drugs to find documents on Twitter and AskAPatient, respectively. This, however, often leads to problems when applying the learned models to new data – they are too adapted to certain drugs or adverse effects mentioned in the training data. Alvaro et al. (2017) attempted to mitigate this phenomenon by creating a corpus (TwiMed) of social media and non-social media sources, annotated with entities based on the same guidelines for both genres to counteract the issue of too many different guidelines for different data, allowing a direct comparison between different genres. Moreover, in cases such as the French SMM4H data, there might be insufficient data for training the (large) LMs nowadays.

Domain-specific Text Variations Further, when working with social media of all kinds, but also with, for instance, reports written by professionals, they contain a lot of genre-typical abbreviations (Segura-Bedmar et al., 2014) and unconventional spellings (Ginn et al., 2014; Segura-Bedmar et al., 2014), lexical variations (Segura-Bedmar et al., 2014), as well as slang or jargon (Huynh et al., 2016), making entity detection and normalization unequally more difficult than when working on scientific texts.

As reported by Sarker and Gonzalez (2015), the short Twitter messages also provide a challenge. In the authors' case, this resulted in not being able to generate a lot of features from these messages, and nowadays, this limits the context Transformer-based models can benefit from. On the other side, some long sequences, as can be found in patient reports, might be much longer than the maximum sequence length required by Transformers.

Incorporating Controlled Vocabularies Ontologies sometimes seem to help and sometimes seem to decrease the performance of the applied methods, depending on how they are used in the process. Segura-Bedmar et al. (2014), for example, report that they introduced a lot of false positive ADR mentions by using an unfiltered MedDRA vocabulary. Metke-Jimenez et al. (2014) show that using Consumer Health Vocabulary (CHV) works better than UMLS for social media, although UMLS is the more comprehensive collection. In more recent approaches for the detection of ADRs, lexicons and ontologies are used less frequently, but with knowledge base integration receiving much attention lately, this might only be a matter of time.

Ambiguities Another factor that makes the detection of drugs and ADRs difficult is the high frequency of ambiguities. For example, often, drug names are similar to women's names (e.g., "Lyrica") or just common words (e.g., "alcohol") also used in other contexts (Segura-Bedmar et al., 2014).

Generic statements are another factor that complicates the detection of relevant documents (Sarker and Gonzalez, 2015). On social media platforms in particular, and especially in times of pandemics, ADRs might rather be rumors and speculations than actual experiences of the writers.

Further, indications and ADRs are often confused by systems, demonstrating that still more context "understanding" is necessary. And finally, as pointed out in the context of n2c2 2018, linking ADRs to their causes is more difficult when several causes are discussed or when long spans of text are written between the cause and the reaction (Uzuner et al., 2020).

3.2.4 Summary

When reviewing the literature on the mentioned tasks with respect to ADRs, it becomes clear that most data is still provided in English, and only a few other languages are represented. In the works described above, this resorts to Japanese, Russian, Spanish, Italian, and French. Some of the data in these languages, however, were never published or are only available without labels.

The methods used do not differ much across languages but often lack the right amount of data. However, even on English, there is much room for improvement in all tasks related to ADR detection. Nevertheless, while in earlier years researchers provided detailed error analyses of the performance of their systems, it is not completely clear what the problems nowadays are, except for the challenges described above, since often, only F_1 scores are reported and nothing else – often, the ADR datasets are simply used as a demonstration corpus for the newest technique and not because the task itself should be improved.

Another problem revealing itself is the fact that datasets from different sub-domains are rarely mixed, and therefore, even for the exact same task in the same umbrella domain, it is difficult to transfer performance from one dataset to the other. But, as put by Chapman et al. already in 2011, the progress in developing NLP techniques for the clinical[14] domain still lags behind the advancements in the general domain. This is, amongst other factors, due to privacy concerns with respect to clinical data and the hence resulting scarce distribution of datasets. Shared tasks like the n2c2, SMM4H, BioNLP, BioCreative, and NTCIR series address some of these problems but still mostly focus on English data. It is, therefore, a task for the international community to improve the situation.

As mentioned at the beginning, most of the above-described work is focused only on tasks related to the detection and extraction of ADRs. Since this involves classification, entity detection, and classification as well as relation extraction, most common general extraction techniques can – with slight adaptions to the domain – probably also be applied here.

Other sub-fields in the biomedical and clinical domain are closely related as well. For example, there is also a huge effort regarding the extraction of medication mentions, chemicals, and other biomedical or clinical entities apart from ADRs. Agrawal et al. (2022), for instance, showed that `InstructGPT` (Ouyang et al., 2022) works well in few-short scenarios for English NER in the clinical domain. Verma et al. (2023) combined the predictions of popular NER systems in the biomedical domain using different strategies and also another model on top and show that this can improve the performance on several widely used datasets.

3.3 Cross-lingual Information Extraction in Biomedical NLP

Some work on non-English data concerned with ADRs was already described in Section 3.2. However, none of these works addressed more than one language at the same time. In their review on "Clinical Natural Language Processing in Languages other than English", Névéol et al. (2018) summarize the non-English publications in the clinical and biomedical domain up until 2017. The majority of these, however, are neither multi- nor cross-lingual.

However, in recent years, thanks to multi-lingual ontologies like UMLS and large multi-lingual models like `XLM-RoBERTa`, the interest in multi-lingual biomedical and clinical NLP increased. Therefore, the below briefly highlights some examples[15]. Note, however, that only two of them are concerned with ADRs.

3.3.1 Datasets

According to Névéol et al. (2018), the CLED-ER evaluation lab in 2013 (Rebholz-Schuhmann et al., 2013) was the first venue that offered a multi-lingual shared task on entity recognition in parallel corpora. In the course of that, one of the first multi-lingual biomedical corpora was

[14] (and biomedical)

[15] The overview is not exhaustive but only emphasizes some works. For publications before 2018, we picked the multi-lingual approaches referred to in the survey of Névéol et al. (2018), the ones after 2018 are picked to represent a variety of sub-domains and tasks in the clinical and biomedical domain.

published: the Mantra corpus (Kors et al., 2015). It contains annotations of entities together with their Concept Unique Identifiers (CUIs).

Neves et al. (2016) published a parallel corpus of biomedical articles from Scielo[16] containing aligned sentences in English, French, Spanish and Portuguese. Further, in 2018, Neveol et al. provided a corpus of death certificates in French, Hungarian, and Italian for the CLEF eHealth 2018 challenge.

3.3.2 Methods & Models

Bodnari et al. (2013) presented a system that can extract biomedical entities in English, French, and Spanish. It is based on a CRF using manually crafted features per language and word alignment methods to transfer annotations between languages.

Duque et al. (2016) attempted to disambiguate medical named entities by using a multilingual approach. They showed a 7% improvement over purely mono-lingual approaches by building a co-occurrence graph from several multi-language datasets that helps to filter candidates for biomedical entities.

For the 2018 CLEF eHealth shared task on ICD-10 coding in death certificates, participants applied techniques such as classification using statistical ML, e.g., random forests, and NNs, e.g., CNNs and RNNs with attention, but also dictionary-based approaches or combinations of both, as well as translations. Seva et al. (2018), for example, built a language-agnostic encoder-decoder approach using multi-lingual `fastText` embeddings to tackle the shared task.

Roller et al. (2018) worked on concept normalization for linking concepts in the French Quaero corpus and the multi-lingual Mantra corpus. The authors show that especially for medical terms in European languages, a simple cross-lingual search improves normalization, which they assumed to be due to the common roots of the words in Greek and Latin.

Hakala and Pyysalo (2019) used `mBERT` combined with a CRF to perform NER in the biomedical domain and show that is works surprisingly well even without being pre-trained on in-domain data. Similarly, Ding et al. (2020) performed cross-lingual transfer-learning for English NER by incorporating Chinese medical data. They pre-trained a bi-lingual `XLM-RoBERTa` model[17], incorporated a medical ontology and further added a transformation matrix that aligns bi-lingual embedding spaces. The token embeddings retrieved from the aligned space were concatenated with the `XLM-RoBERTa` embeddings and fed to a bi-LSTM with a final CRF layer. The adapted model gains about 3 points in F_1 score over the original `XLM-RoBERTa` on the English i2b2 2010 dataset. Mutuvi et al. (2020) experimented with the multi-lingual classification of epidemiological datasets. They showed that `mBERT` outperforms all baselines using "traditional" ML and neural architectures.

For the SMM4H 2020 shared task, Miftahutdinov et al. (2020) compared different model and data setups and showed that a CNN in combination with `fastText` can outperform `mBERT` on Russian ADR texts in binary classification when the `fastText` embeddings were trained on Russian health-related data. When using both English and Russian tweets for fine-tuning an English-Russian `BERT` model (`EnRuDRBERT`) and using an ensembling approach, they achieved the best result (within their experiments), especially when compared to only fine-tuning on Russian data. However, the authors also note that adding Russian data to the English data only improved the results on the English test set by one percentage point.

Raval et al. (2021) investigated the performance of `mBERT` and `T5` trained on English when applied in zero-shot mode to the French SMMM4H 2020 data. They showed that their setup

[16] `https://scielo.org/`

[17] It is not entirely clear if the authors used the `XLM` or the `XLM-RoBERTa` model based on their descriptions of the language modeling objectives.

using T5 (Raffel et al., 2020) with proportional mixing and temperature scaling allows a zero-shot transfer to the French imbalanced data which results in an F_1 score of 20%, better than when using mBERT and the same result as the best system in SMM4H 2020 achieved.

Frei and Kramer (2022) took English n2c2 2018 data (Henry et al., 2020) and automatically translated and aligned it to German using the fairseq (Ott et al., 2019) and fastAlign (Dyer et al., 2013) frameworks, creating a German n2c2 dataset. They showed that a model trained and tested on the newly created corpus achieves an F_1 score of approximately 81%. A similar approach was followed by Schäfer et al. (2022) on the German BRONCO corpus (Kittner et al., 2021) and three English corpora.

For clinical NER, Gaschi et al. (2023) compared, again, translation of training or test set with cross-lingual transfer. They used the dataset provided by Frei and Kramer (2022) and released a new French dataset intended as an evaluation dataset. They show that translation-based approaches can be similarly successful than cross-lingual methods, but require more effort and careful design.

Finally, Meoni et al. (2023) study multi-lingual medical entity extraction using large LMs, similar to the work of Agrawal et al. (2022) who also use InstructGPT (Ouyang et al., 2022). In contrast to them, however, Meoni et al. (2023) use Large Language Models (LLMs) to pre-annotate EHRs with which local confidential models can be trained in the clinical domain.

3.3.3 Summary

Most approaches in cross- and multi-lingual IE in biomedical NLP are very similar to those conducted in the general domain on multi-lingual texts. Especially translation approaches and, nowadays, multi-lingual models are popular means, often also in combination, for compensating the low amount of target language data and/or annotations. There is not a huge body of related work right now, but it seems to be growing, since now, researchers have the means to get more successful results, and institutions recognize the value multi-lingual information access can provide for patients and medical professionals.

3.4 Existing Datasets

As mentioned before, there are quite a few annotated datasets tackling questions in biomedical and clinical NLP by now, although not multi-lingual ones.[18] The number of supported languages has also been growing in recent years. On the Huggingface data hub[19], at least six languages for health-related datasets are represented: English, Spanish, Chinese, French, German, and Japanese. Of course, this is not a lot, but at least a beginning, and also, the list is not exhaustive since some datasets cannot be easily shared on a hub like this due to data privacy concerns.

Unfortunately, only a few datasets include annotations for the detection and extraction of ADRs. Although NLP in the domain of pharmacovigilance has been researched for quite some time, usable, that is, publicly available annotated data is still scarce, particularly for languages other than English. Further, only a few of these datasets represent a "reversed" perspective, namely the one of the patient. Indeed, most data sources are written by experts in the field, either practitioners or scientists, who write reports or papers about specific cases.

Since approximately 2010, with one of the first publications on the extraction of ADRs by Leaman et al. (2010), the interest in and the number of social media datasets has been growing slowly since researchers, health-related industries, and authorities recognize the value of

[18] At the time of writing, the best dataset overviews are probably given on huggingface spaces, where the BigScience sub-group for biomedical NLP aims to collect all publicly available datasets, as well as a list of datasets provided by Fries et al. (2022), which seems to be an intermediate version (https://tinyurl.com/bigbio22).

[19] https://huggingface.co/bigbio

patient-generated data with respect to improving medication products and public health monitoring (Sarker et al., 2015). For English social media datasets published between 2010 and 2014, we refer the reader to Table 1 in the review of Sarker et al. (2015). Unfortunately, not all of these are publicly available.

An overview of publicly available *non-English datasets* is provided in Table 3.1. The listed datasets and their annotation process are described in the following to highlight their differences and emphasize the challenges associated with each corpus, especially in comparison with the new corpus provided by this work.

lang.	#docs	neg	pos	ratio	type	annotation	authors
es	400	235	165	1.4 : 2	forum	entities	Segura-Bedmar et al. (2014)
fr	3033	2984	49	61 : 1	Twitter	binary	Klein et al. (2020)
ru	-	-	279	-	drug reviews	multi-label	Alimova et al. (2017)
ru	*500	-	-	-	drug reviews	multi-label + entities	Tutubalina et al. (2020)
ru	9515	8683	842	10 : 1	Twitter	binary	Klein et al. (2020)
ja	169	-	-	-	forum	entities +normalization	Arase et al. (2020)

Table 3.1: Other non-English social media corpora for the detection (and sometimes extraction) of ADRs. fr=French, ru=Russian, ja=Japanese, es=Spanish. The number of documents (#docs) refers to the definition of documents per corpus, i.e., some are sentence-based, some are post-based, etc. The annotations were converted to binary classes where possible. Some test sets are unavailable to the public since they are/were part of a shared task. *This is only the annotated part of the RuDReC corpus

Spanish (es): The SpanishADR corpus (Segura-Bedmar et al., 2014) was the first non-English social media dataset focused on ADRs overall. The data they collected was downloaded from a Spanish patient forum called "ForumClinic"[20]. The authors downloaded all available data at that point and randomly picked 400 forum posts for annotation. Then, two annotators "with expertise in Pharmacovigilance" (Segura-Bedmar et al., 2014) annotated Adverse *Events*[21] and drug mentions, where drug mentions refer to "substance[s] used in the treatment, cure, prevention or diagnosis of diseases", using an annotation tool provided by the open source software toolkit GATE (Cunningham et al., 2013). Both generic and brand names of medications as well as medication groups were annotated. Also, mentions with errors (grammatical or spelling errors) and nominal anaphoric expressions were included. The work of both annotators was merged with the help of a third annotator to solve disagreements. Segura-Bedmar et al. (2014) report an IAA based on F_1 score of 0.89 for the drug mentions and 0.59 for Adverse Events (AEs).

Japanese (ja): Similar to the Spanish corpus, Arase et al. published a corpus in 2020 based on the Japanese patient forum called TOBYO[22]. The authors crawled all entries related to lung cancer and containing one to five drugs out of a pre-compiled dictionary. They further filtered

[20] http://www.forumclinic.org

[21] An adverse event is "any untoward medical occurrence in a patient to whom a medicinal product is administered and which does not necessarily have a causal relationship with this treatment" (https://learning.eupati.eu/mod/book/tool/print/index.php?id=811).

[22] https://www.tobyo.jp/

the data and sampled 500 posts randomly for annotation. The final corpus provides annotations of drug effect spans, related drug mentions, types of reactions, and the ICD-10 codes for those. The entities were labeled with IOB tags on character level.

The data was annotated with the help of trained annotators, who were, however, no experts in the fields of medicine or pharmacy. Microsoft Word was used for annotation. The annotators identified drug names and adverse reactions to those medications and labeled them with the respective markers, such as ICD-10 codes for symptoms, including both negative and positive effects of the drug. General expressions for drug names, like "tablet", were excluded, but brand names, as well as specific medical substances, were annotated. Drug effects that are described by patients but not actually experienced by them were *not* annotated, neither were symptoms not related to a drug.

The identified effects were further divided into target and adverse reactions using four different labels. "Target-effect positive" describes the intended effect of the drug that actually eventuated. In contrast to that, "target-effect negative" is the label for desired effects which *did not* occur. "Adverse-effect positive" is an ADR which is known and happened to the patient, while "adverse-effect negative" is an ADR which should have happened, but did not. Finally, annotators were asked to assign ICD-10 codes to each effect by querying MANBYO-SEARCH, a search engine that receives a reaction expression and returns one or more possible codes.

The work of all three annotators was consolidated by taking the label set that at least two annotators chose. This means the annotators had to agree on the drug name, IDC-10 code, and effect type. Sets of labels produced by only one of the annotators, and therewith the entire article, were discarded. Then, the longest provided span was selected. This process resulted in 169 articles out of 500. Arase et al. (2020) explain that most of the discarded data contained information irrelevant to ADRs. However, with this, the authors remove all negative examples of documents containing ADRs. Inter-annotator agreement was calculated using Fleiss' κ, resulting in $\kappa = 0.52$ for span and type agreement.

Russian (ru): **Alimova et al. (2017)** provided the first dataset of this kind for the Russian language. They crawled the drug review forum Otzovik[23] and created a corpus based on 580 reviews with respect to specific medication, e.g., antiviral and soporific drugs. The annotators were "specialists in the field of medicine" (Alimova et al., 2017) and the annotation instructions were based on the work of Leaman et al. (2010). The reviews were split into sentences using the Texterra system (Turdakov et al., 2014) and marked with one out of four labels: *Indication, Beneficial effect, Adverse drug reaction, Other*. Sentences that could be annotated with more than one label were removed (69 sentences), and therefore, sentences with the label *Indication* only contained the *reasons* for taking a specific drug, i.e., symptoms and diseases (646 sentences in total). *Beneficial effect*, as the name suggests, is the label for sentences describing only recovery reports of patients (335 sentences). *Adverse drug reaction* marks sentences as containing a description of a decline in a patient's health (279 sentences). Finally, *Other* describes all cases that do not fit into the other three labels, resulting in 4,488 sentences. The authors did not state how many annotators worked on the data and neither reported the IAA.

Tutubalina et al. (2020) created another Russian dataset of drug reviews three years later, calling it the Russian Drug Reaction Corpus (RuDReC). The data is divided into two parts: one containing annotations on entity level, the other one without annotations. The annotated corpus is based on online drug reviews from the same forum that was used by Alimova et al. (2017) and comprises 500 documents. The second part of the corpus contains, according to the authors, 1.4 million texts from various online sources focused on health-related user-generated posts.

[23] http://otzovik.com

The labels of the annotated corpus are sentence-based and mark the existence or absence of health-related issues using five different sentence labels. Those that contain health problems were further annotated on entity level, distinguishing six different entity types. According to Tutubalina et al. (2020), the annotation guidelines are in line with those of Karimi et al. (2015) and Zolnoori et al. (2019).

Annotation was carried out in two steps using INCEpTION (Klie et al., 2018). During the first step, four annotators – all with a background in pharmaceutical sciences – were asked to highlight relevant spans of text in 400 test examples. These spans included drug names and patients' health conditions at various points in time and related to drug intake. Using the pre-annotations of the first annotation round and discussions with the annotators, IAA was determined to be "approximately 70%", using a relaxed agreement for disease and drug entities following earlier work (Karimi et al., 2015; Metke-Jimenez et al., 2014). Based on this, the following sentence and entity labels were selected in the end. For sentence-level annotation, Tutubalina et al. (2020) decided on *DE* (drug effectiveness, the sentence contains a report about an improvement of the patient's health or about treated symptoms), *DIE* (drug ineffectiveness, the health of the patient deteriorated or the medication did not have any effect), *DI* (the sentence described the reason, e.g., symptoms, of medication intake), *ADR* (the sentence contains a report on undesired reactions due to medication intake) and finally, *FINDING* (events related to diseases not experienced by the patient, e.g., absence of drug effects). Entities are either labeled as *DRUGNAME* (brand names or product ingredients), *DRUGCLASS* (general mentions of drug families), *DRUGFORM* (routes of medication intake), *DI* (drug indications and symptoms), *ADR* (negative events occurring as a consequence of medication intake, not associated with the symptoms), or *FINDING* (*DIs* or *ADRs* not directly experienced by the patient, entities about which the annotator was not clear).

For the second step, two annotators continued the annotation process with the determined sentence and entity labels. After finishing, their work was reviewed by the authors. The dataset contains reviews of four different groups of drugs. It is not entirely clear if those drugs were targeted in the first place or if they were only emerging after annotation.

Klein et al. (2020) present a Twitter dataset made from Russian tweets with binary annotation. The training set (which is the only one available) contains 7,612 tweets of which 666 describe an ADR. For the test set, Klein et al. (2020) list 1,903 tweets with 166 containing an ADR. The tweets were first annotated by three annotators from Yandex Toloka[24], then consolidated into one single label, and finally reviewed by an additional annotator. The authors do not mention the background of the annotators but report an IAA of 0.49 using Cohen's κ. The data was presented for the fifth Social Media Mining for Health Applications (SMM4H) shared task in 2020.

French (fr): The French corpus (Klein et al., 2020) is based on data collected from Twitter as well and was also part of the SMM4H 2020 shared task. The training set contains 2,426 tweets with 39 ADR examples, while the test set (which is not publicly available) comprises 607 tweets, with only 10 texts describing an ADR. 848 tweets were double annotated by three annotators (no background given) using binary labels. On these, the authors report an IAA of 0.61 and 0.69 for each annotator pair.

English (en): Since we will make use of two English datasets in Chapter 5 as well, we will briefly present those, too. Both datasets are based on the patient forum AskAPatient[25].

Karimi et al. (2015): The CSIRO Adverse Drug Event Corpus (CADEC) is the most widely known dataset for the detection of ADRs from social media. In total, it contains 7,398 sentences

[24] https://toloka.yandex.ru/

[25] https://www.askapatient.com/

from 1,253 user posts. The posts were provided (and not crawled) by the AskAPatient forum based on 12 pre-defined drugs, for example, *Cataflam* or *Arthrotec*. Each of those drugs belongs to either the group of *Diclofenac* or *Lipitor*. They were annotated in two steps using the annotation tool BRAT by medical students and computer scientists and finally screened by three of the authors to correct mistakes. During normalization, the annotations were further reviewed by a clinical terminologist (Karimi et al., 2015).

Annotators were first asked to identify relevant entities, which were then linked to a controlled vocabulary, in this case, SNOMED-CT[26], Australian Medicines Terminology (AMT), and MedDRA, depending on the entity. Also, annotation was executed on the sentence level, while entities crossing sentence boundaries were omitted. However, entities were allowed to be discontinuous but not embedded (an example of a discontinuous entity is shown in Section 4.1.3). Also, generic mentions, like "side effect", were not annotated. This is also true for co-references or anaphoric expressions. Furthermore, spans were limited to not include prepositions, qualifiers, or possessive adjectives.

The entities to be annotated were finalized in consultation with annotators and experts in the field as follows: *Drug, ADR, Disease, Symptom*, and *Finding*. Most of these are self-explanatory, *Finding* represents, similar to the work of Tutubalina et al. (2020), events not directly experienced by the patient or other occurrences about which the annotator is not clear.

The actual annotation started after an initial pilot annotation task and adapting the annotation setting accordingly. According to Karimi et al. (2015), the forum posts were given in equal parts to the annotators, with an overlap of 55 documents for the calculation of IAA. The authors calculated strict and relaxed agreement, following the work of Schouten (1980); Metke-Jimenez et al. (2014). In a relaxed setting, annotations for the *Diclofenac* group achieved an IAA of 78% (four annotators) while the *Lipitor* group resulted in an IAA of 95% (two annotators).

In the second stage of annotation, the normalization, only one annotator, a clinical terminologist, was working on the data. All entities except *Drug* were mapped to Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT), mostly in a one-to-one fashion but sometimes also in a one-to-many fashion when it seemed necessary.

All entities labeled with *Drug* were linked to entries in AMT. Here, the authors had a similar problem as Segura-Bedmar et al. (2014): Since AMT is a collection of medication released in Australia, but AskAPatient can be used by English-speaking people around the world, some drug names were not found in the terminology. In those cases, a generic concept or "concept_less" as a dummy was annotated. Finally, ADRs, or rather their concepts inferred from SNOMED-CT, were mapped to MedDRA, specifically to the Lowest Level Term (LLT) to cope with the colloquial language.

Zolnoori et al. (2019) created a similar corpus, but with a focus on psychiatric medications. The Psychiatric Treatment Adverse Reactions (PSYTAR) corpus provides three types of annotation: sentence-level labels, entity annotations and normalization. The forum posts crawled from AskAPatient all contain one out of four psychiatric medications: *Zoloft* and *Lexapro* (from the class of Selective Serotonin Reuptake Inhibitors) and *Cymbalta* and *Effexor* (from the class of Serotonin Norepinephrine Reuptake Inhibitors). They were pre-processed by using regular expressions to replace personal information and noisy patterns, for instance, errors in punctuation.

The authors sampled a number of posts from the downloaded data and split them into sentences. They were first annotated on sentence level for the occurrence of ADRs, withdrawal symptoms, general signs or symptoms, drug indications, drug effectiveness, and drug ineffectiveness. If a sentence did not report any of the aforementioned events, it was labeled as *Others*.

[26] See Appendix A.4.

In a second step, spans describing ADRs, withdrawal symptoms, general signs or symptoms, and drug indications were annotated. Also, qualifiers with respect to the entity types were annotated. The guidelines were based on earlier work of the authors (Zolnoori et al., 2017). For example, if a patient was not certain about the connection between medication intake and an adverse effect, this was *not* annotated as an ADR. However, subjective complaints, functional problems due to medication intake, and duplicated mentions were extracted. In contrast, constructions like metaphors or figures of speech were omitted.

Lastly, the annotators normalized these entities using UMLS and SNOMED-CT and categorized them into the classes *physiological, psychological, cognitive*, and *functional problems*. For normalization, if the annotators could not find a fitting concept following a flow chart provided by the authors, in neither UMLS nor SNOMED-CT, the entity was assigned to the dummy concept "no-codes".

For sentence classification and entity annotation, the authors hired four annotators with either a background in pharmaceutics or health sciences. Each review post was double-annotated, and IAA was calculated based on Cohen's κ. The IAA for sentence classification resulted in a κ of 0.78. Sentences on which the annotators disagreed were revisited and resolved by the same annotators. Disagreement on entities was resolved by one of the authors, and IAA was measured using pair-wise agreement (Schouten, 1980). It resulted in a score of 0.86 for the entire dataset, 0.86 for the agreement on ADRs, 0.81 for withdrawal symptoms, 0.91 for signs and symptoms, and 0.91 for indications.

Differences in the described corpora Additionally to the n-ary format on document-level, some of the corpora also offer a more fine-grained annotation and, therefore, more detailed information. This includes the annotation of entities but also normalizing entities to their medical concepts from various ontologies, for instance, SNOMED-CT and UMLS. All works described above tackle various aims with the creation of their corpora, resulting in different annotation schemes for different use cases. This resulted in a different number of document and entity labels per corpus, mostly depending on granularity and focus. More "complicated" documents, like those associated with more than one label or ambiguous ones, are discarded in some works (Alimova et al., 2017; Arase et al., 2020).

The pre-selection of medications also plays an important role. Some authors, e.g., Arase et al. (2020), focused on specific diseases, and the associated medication, some chose a certain set of drugs (mostly) independent of the disease (Karimi et al., 2015; Zolnoori et al., 2019; Klein et al., 2020), some just took everything they could get and sampled randomly (Segura-Bedmar et al., 2014). For example, the two most similar corpora, CADEC and PSYTAR, are different in that they focus on non-overlapping medication (groups) and have a different granularity of annotation (e.g., review-level versus sentence-level). Further, entities in CADEC were mapped to MedDRA, AMT, and SNOMED-CT, while those of PSYTAR were mapped to UMLS. In contrast, the Japanese dataset created by Arase et al. (2020), for example, is based on paragraphs, and diseases, signs, and symptoms are mapped to ICD-10 codes. However, medication names, although annotated, are not associated with any ontology.

Some of the corpora include annotations of drug ineffectiveness (Zolnoori et al., 2019; Tutubalina et al., 2020; Arase et al., 2020), while the others do not have annotations of these entity spans. Sometimes, generic expressions like "side effect" or "pill" are included (Segura-Bedmar et al., 2014) while in other datasets, these are deliberately excluded (Arase et al., 2020; Karimi et al., 2015). The same applies to anaphoric and other referencing expressions, modifiers, possessive pronouns, and qualifiers. Some datasets include them (e.g., anaphoric expressions are annotated by Arase et al. (2020)), others do not (e.g., anaphoric expressions are excluded by Karimi et al. (2015) by design). Note that some authors do not give information about how they handle these expressions.

One noticeable difference of PsyTAR (Zolnoori et al., 2019) compared to all other corpora is the inclusion of what they call "functional problems", meaning issues that negatively affect the patients' social life, daily functioning and, in general, quality of life. The authors included these because the psychiatric medications they focused on often have an influence on the aforementioned areas.

The length of examples also varies depending on the underlying source. Twitter messages are, in general, rather short, while drug reviews and forum posts are longer. However, some of the authors intentionally shortened (Arase et al., 2020) or split up (Zolnoori et al., 2019; Tutubalina et al., 2020) the documents. Moreover, depending on document length, some authors (Karimi et al., 2015) allowed cross-sentence annotations as well as discontinuous entities (Karimi et al., 2015). However, most authors did not comment on these phenomena. Note that none of the presented corpora, however, has marked any relations between the entities. Karimi et al. (2015) mentioned being "in the process" of adding relations and more entities, but this data, if existent, is not available.

Finally, but also importantly, we see that the calculation of IAA is done differently in almost all corpora or not at all (Alimova et al., 2017). Segura-Bedmar et al. (2014) report using F_1 score for the IAA of entities, where one out of two annotated datasets is used as the gold standard. Arase et al. (2020) used Fleiss' κ to calculate agreement on entities. Tutubalina et al. (2020) follows previous work (Metke-Jimenez et al., 2014; Karimi et al., 2015) and report pairwise agreement (relaxed) on entity level. Karimi et al. (2015) evaluated the annotated entities using both strict and relaxed pairwise agreement. Zolnoori et al. (2019) use Cohen's κ for sentence-based annotations and pairwise agreement for evaluating the entity-level annotations. Both datasets of Klein et al. (2020) were evaluated using Cohen's κ on the sentence labels.

All of this makes it rather difficult to compare datasets and experimental results using these datasets across languages but even within a language. The IAA, in particular, is an important means to evaluate and validate the quality of a dataset and should be calculated in a consistent way or, if possible, from multiple perspectives, i.e., using several measures.

As with other sub-domains in biomedical NLP, common annotation schemes and guidelines are important to compare performances and to transfer models and insights. Of course, if a dataset focuses on a specific domain, then entities specific to this domain have to be annotated. Nevertheless, annotating similar datasets using a similar or even the same annotation scheme might also provide interesting insights specific to the languages and cultures. Also, adding some annotations that might be interesting for another task or domain might not provide too much overhead and can be beneficial in unsuspected ways.

Going back to the presented corpora, in particular the non-English ones, it becomes clear that there is still a need for more data in both the already tackled languages since the datasets are rather small, but also in "new" languages, to improve the monitoring of public health. Of course, under-resourced languages should be represented as well, but in this case, even languages that usually have enough resources are not covered, for example, French and German.

3.5 User Privacy

When it comes to biomedical text processing on social media, researchers also have to deal with the privacy appertaining social media users. The ethics of working with (data of) social media users have been a widely discussed topic for quite some time (Grimmelmann, 2017; Ford et al., 2021). This does not only concern the written texts of users but, more generally, any personal information they share, sometimes unwittingly. Information that fall under this category might contain their gender, age, geo-location, preferences of any kind, and (network) connections to other persons. Since this work is about text processing, we will focus on the elicitation of tweets, Facebook posts, and other social media messages that are relevant to works in biomedical NLP

and the problems that arise when using and distributing these messages. For the latter, we will summarize the findings of other research regarding social media mining for health. We also draw heavily on a recently published review by Ford et al. (2021).

3.5.1 Data Types and Collection Methods

With respect to social media platforms, Twitter, in particular, plays an important role. Tweets used to be easy to collect since they could be downloaded without cost[27] using Twitter's official streaming API[28] (Nikfarjam et al., 2015b; Paul et al., 2016; Sarker, 2017). After downloading, the collected messages were usually annotated according to the tasks they were collected for, and the annotations were made publicly available, either via publications or during shared task challenges (Sarker and Gonzalez, 2015; Klein et al., 2017; Weissenbacher et al., 2018, 2019, and others). If the data were not shared directly, authors often provided tweet IDs and a downloading script such that challenge participants or other researchers could collect the tweets by themselves (Weissenbacher et al., 2018, 2019). Although this is a good way to circumvent sharing the tweets and being responsible for keeping them secure (with this strategy, every participant is responsible on their own), it often happened that after a while, not all of the tweets were downloadable anymore, because Twitter users might have deleted them. , This is not only unfortunate for the annotation work that went into the data since it was now in vain, but also reduces the dataset size. That, in turn, might impact the performance of ML models that are supposed to be trained on these data, especially when the data's labels were imbalanced to begin with. Also, the comparability and reproducibility of experiments suffer.

However, there is no way to ask Twitter users directly whether they agree with their data being shared. Especially when thousands of tweets are needed for research, it is simply not feasible to reach out to the original authors of the messages. Therefore, researchers often add a statement to their publication, declaring that their work was approved by an official review board or ethics committee (Weissenbacher et al., 2019; Basaldella et al., 2020), based on the fact that tweets (and other social media posts) are accessible by everyone on the internet.

Apart from Twitter, another popular choice for getting user-generated texts associated with health topics is Reddit. Reddit provides so-called "sub-reddits" for users to interact with each other within a dedicated topic. Sub-reddits can be very general but also very specific, e.g. discussing a particular medication only. Thus, in contrast to Twitter, it is easier to find relevant data for a specific research question. Basaldella et al. (2020), for example, used Reddit to build a corpus for medical entity linking. They argued that Reddit is both anonymous and publicly available. Further, they also de-identified the Reddit posts "as far as possible" to keep the users' privacy (Basaldella et al., 2020), meaning that they removed identifying information such as names, user handles, e-mail addresses, and so on. Another Reddit corpus was created by Lavertu and Altman (2019). They downloaded the data using an API[29] for social media data dumps. However, it is not clear in which way this API is associated with Reddit. The authors refer to the "pseudo-anonymous nature" of the data they are using (Lavertu and Altman, 2019), but they did not mention any procedures of de-identification. Finally, a third Reddit corpus resulted from the work of Scepanovic et al. (2020). The authors do not describe the exact procedure they used for collecting the data but only state that they downloaded the data from Reddit. They further provided the data to Amazon Mechanical Turk workers for annotation (Scepanovic et al., 2020) and therefore released data to people not even in the research community and possibly without any data protection agreement. Note that these three publications are not the only works that use Reddit data; they serve as examples.

[27] As of the time of writing this thesis, Twitter is not accessible anymore (and neither is it called Twitter anymore).
[28] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data
[29] https://pushshift.io/

The third popular choice of biomedical text dataset sources are patient and medication review forums. In contrast to tweets, and more similar to Reddit, forum messages tend to be much longer and provide more context. On the other side, they also can be *too* long, covering a lot of different topics and diverging from the original post, which distinguishes them from the Reddit posts. Segura-Bedmar et al. (2014), for example, were the first to collect a corpus for Spanish health-related texts from the forum ForumClinic[30], which was discussed in Section 3.4. They argue that the user posts are publicly available, and so they used a web crawler to collect as many documents as were needed for their study (Segura-Bedmar et al., 2014).

Two other datasets based on the English "AskaPatient" drug review forum[31] were discussed in Section 3.4 as well. Karimi et al. (2015) reported that the data was provided by AskaPatient themselves, and the project was approved by the forum's ethics committee. It is downloadable by agreeing to a CSIRO Data License[32]. They did not mention any de-identification process. Zolnoori et al. (2019) used the same forum as data source but a different set of drugs than Karimi et al. (2015) to filter users' posts. Regarding the collection process, they argued that the data is publicly available and used a web crawler to get the drug reviews they were interested in. They de-identified the data using regular expressions and distributed the corpus under a CCBY 4.0 Data license. It is, therefore, freely available, too.

There is also the possibility of synthesizing data. This can be done by either (back-)translating data that was already de-identified (Frei and Kramer, 2022, inter alia) or by generating completely new data by using generative language models like T5 (Raffel et al., 2020). For example, generated data like this is provided for the MedNLP-SC shared task in 2023[33] which will be described in Section 4.2.2. However, so far, these generated data are still not the same as original tweets and lack, for example, the "Twitter style" of writing and sometimes also textual coherence. Nevertheless, this might be an exciting new avenue for preserving user privacy in biomedical NLP.

3.5.2 Ethical Issues

Collecting data from social media for biomedical or clinical NLP, as described above, is accompanied by several ethical issues. First of all, users whose data are collected usually do not know about their messages being used for analysis or as training material. Even though some users might even think of having "given up" their right to privacy when registering for such a platform (Ford et al., 2021; Mikal et al., 2016), it *still* hurts their privacy.

This is followed by the question of ownership: Who does the text belong to after its publication on a platform? According to Ford et al. (2021), many researchers argue that publishing social media posts is equivalent to giving consent to the use of the content for any purpose. Thus, they believe that the content belongs to "the public", which is, in the very least, a debatable proposition, but one that again and again divides the research community (McKee, 2013; Paul et al., 2016).

Ford et al. (2021) further describe the various data policies executed by different websites: for some, users can restrict their posts to certain groups of "friends" or topic-dedicated subgroups (for instance, Facebook or Reddit), while others do not have any restrictions on the audience, for example, Twitter, where even people not registered on the platform could read almost everything.

Another issue arising from the former is that even if something was published in a public space, it still does not mean that it can be *re-used* by third parties (Ford et al., 2021). Some

[30] http://www.forumclinic.org/
[31] https://www.askapatient.com/
[32] https://confluence.csiro.au/display/dap/CSIRO+Data+Licence
[33] https://sociocom.naist.jp/mednlp-sc/

platforms, however, already integrate third-party use in their data protection agreement, for example, Twitter[34].

As further mentioned by Ford et al. (2021) and others (Moreno et al., 2012; Mikal et al., 2016), users often lack knowledge when it comes to the privacy settings of social media platforms, which are often difficult to understand for laypersons (Mikal et al., 2016). Moreover, users might underestimate the ways in which their data might be re-distributed and used (Ford et al., 2021), following the notion of being "unimportant" (Mikal et al., 2016). They also often cannot oversee how far back their data is still accessible (Ford et al., 2021).

In conclusion, Ford et al. (2021) argue that even though social media data might be used for research if properly de-identified, researchers should still follow ethical guidelines (e.g., by mitigating against potential harms) and aim for transparency regarding their methods and purpose of research.

Similarly, a recent review by Bear Don't Walk et al. (2022) investigates the state of biomedical text processing through an "ethic lens" (Bear Don't Walk et al., 2022), particularly in context with biases and fairness. They argue that while biomedical text processing might help to advance population health, it is also prone to either enforcing already existing or introducing new biases. This is often due to the use of "black-box" machine learning models, but also based on participant group representation, biases in data collection or documentation practices, and various other factors (Bear Don't Walk et al., 2022). Although in their review, the authors describe the case of applying biomedical text processing methods to clinical text, this is also something to be kept in mind when handling social media data, maybe even more so, since as mentioned, social media users often do not even know their data was used.

[34] https://bit.ly/3rwK5A0

Chapter 4

Data

The entire foundation of today's success of neural LMs is data. The common assumption until recently was "the larger the model, the better the performance" (see Section 2.4.5), but for large models, huge amounts of data are needed.[1] For example, for `XLM-RoBERTa`, about 2.5 TB of data were collected. Despite recent efforts in low- or zero-resource learning (Brown et al., 2020, inter alia), a majority of research in NLP is still driven by labeled data, in particular when it comes to more specific downstream tasks. Also when setting aside data-hungry LMs, a representative amount of data is necessary to draw meaningful conclusions from them, for example for evaluation. Also, the quality and diversity of text and annotations play a key role.

However, both domain-specific texts and annotations are hard to come by. Usually, data simply crawled from the internet, either from specific sources or in general, are of mixed quality and can only be filtered heuristically (Raffel et al., 2020, for example).[2] Also, with the success of generative models like the GPT series (e.g., GPT-3 (Brown et al., 2020)), supposedly user-generated texts found on the internet might not even be human-created. For specialized domains in particular, it is therefore still necessary to collect and annotate appropriate data and ensure that these data are of high quality. Considering this, and the fact that there is no data available for the detection of ADRs in French, German, or Japanese user-generated texts, Section 4.1 describes the process of creating a new multi-lingual corpus for the domain of pharmacovigilance, including data collection and de-identification (Section 4.1.1), guideline development (Example 4.5 and Section 4.1.3), and subsequent annotation.

Another issue, which arises specifically in the field of health-related NLP, is the privacy of patients and their data. Although there exist a variety of public fora, online patient associations, and drug review platforms, and people talk about their health on social media like Reddit and Twitter, actually *accessing* and *using* these data is much more complicated than it seems. *Publishing* these data, annotated or not, to provide other researchers with reproducible materials, is even more complicated. These issues and potential solutions are described in Section 4.2.

Most of the presented work (except for Section 4.2.1) was developed within the tri-lateral project KEEPHA[3] and thus, the three majority languages of the involved countries, German, French and Japanese, are the focus going forward.

4.1 Development of a Multi-Lingual ADR Corpus

This part of the thesis describes the process of developing a multi-lingual corpus of UGT for the classification and extraction of ADRs. It is dedicated to answer RQ 1. In the course of this section, the development of a corpus of texts in French, German, and Japanese with the following levels of annotations is laid out:

[1] And of course, also huge amounts of computing power, which we might run out of in the future.

[2] Note that *quality* in this context does not refer to typos, bad style or grammar, since these characteristics are representative of human writing, but rather to e.g., code published on websites, URLs, boilerplate text and code etc.

[3] Knowledge-Enhanced information Extraction across languages for PHArmacovigilance, `https://keepha.lisn.upsaclay.fr/wiki/doku.php?id=project`

1. Binary annotation: does the document contain an ADR or not?
2. Annotation of entities: adding entity types to mentions of drugs or symptoms and other relevant expressions.
3. Annotation of relationships between these entities, for example, to express that a drug caused a symptom.

We will start by describing the first step, the collection of data, in Section 4.1.1, followed by the development of the cross-lingual guidelines in Example 4.5. Then, the actual annotation process is presented, including the resulting IAA. Finally, we end with the limitations of the corpus as of now (Section 4.1.4), the future plans on how to improve it further, and the lessons learned in Section 4.1.5.

Note that all three languages considered for the now presented corpus descend from different language families, where German belongs to the Germanic languages, French to the Romance languages, and Japanese to the Japonic (or Japanese-Ryukyuan) language family.

4.1.1 Data Collection

Various data types were already described in the preceding chapter. Since we would like to change perspective and take a look at the view from the patient's side, social media presents itself as an ideal type of data. In recent years, social media use has increased and more platforms are used to discuss and share. Currently, popular platforms are, for example, Reddit, Twitter, or Facebook[4].

For our work, we include health-related discussion and review fora into the domain of "social media", since they provide the view from a layperson's perspective where users employ their own language and descriptions.

User forums generally have a main topic but different sub-discussions (threads). For example, the German med2 forum[5] is, according to the website, for "discussions about medical topics, problems in everyday life or health limitations". Thus, some of the threads are related to health issues, but some are just about everyday topics or even political or social discussions. Most of these forums are administrated by volunteers. There are also some professionally administrated fora, like AskAPatient[6], operated by the Consumer Health Resource Group.

Each team in the KEEPHA project was responsible for acquiring the respective data.[7] The general requirements we set for the data were as follows:

1. The data should be *health-related*, but not specific to any drug or disease.
2. The data should be *de-identifiable* or already de-identified.
3. The data should be *distributable* to other research teams.

Requirement 1 made sure that we did not aggregate user posts related to only one drug or disease. In contrast, we wanted the data to be as "natural" as possible, representing the true distribution of ADRs. Also, we wanted to verify whether the amount of ADR mentions in non-English UGTs is comparable to the percentages provided by the literature. Requirement 2 was set to make sure that the posts we collected were already from anonymous sources. This makes it easier to de-identify any remaining Protected Health Information (PHI), i.e., information which are private to the patient. Requirement 3 was the most challenging one as we discovered.

[4] www.reddit.com, https://twitter.com, www.facebook.com
[5] https://med2-forum.de/
[6] https://www.askapatient.com/
[7] The author was part of both the German and French teams.

We wanted the data to be shareable (with a data protection agreement) to be able to make our research reproducible. If a dataset cannot be used by other researchers, they cannot analyze it themselves, benchmark their models or test their hypotheses in general. Therefore, getting data that are distributable was a crucial point in the acquisition process.

Japanese

According to the requirements specified above, the Japanese texts were collected from both Twitter and Yahoo! JAPAN Chiebukuro[8], a Japanese health Q&A forum. Note that for Twitter, we had to relax Requirement 1, since searching for tweets without keywords is not possible. The Japanese team organized the collection, annotation, and analysis of the Japanese data. If there were any disagreements or problems with respect to the annotation guidelines (described below) for one language, we discussed them in the monthly KEEPHA meetings or more often, if necessary. Since the Japanese team created the Japanese data (but with the same guidelines), the details of this part of the dataset are not discussed in this thesis.

German

For the German data, we first contacted a multi-lingual drug review forum where users reported their experience with whatever medication they were taking. This forum seemed promising since it already contained drug reviews in several languages, French, German, and English among them, and accompanying labels for the existence and strength of ADRs. However, after several rounds of discussion with the platform owners trying to come up with an acceptable Non-Disclosure Agreement for both sides, it became clear that the company providing this forum did not want us to share the data with other researchers, even when de-identified and protected by a data protection agreement. This detour, which took over one year, demonstrates the complexity of gathering health-related data for research.

Parallelly, we continued the search and finally got permission from the administrators of the fora lifeline.de[9], henceforth Lifeline, to download and share the data. Lifeline is an independent health platform of the FUNKE DIGITAL GmbH[10], financed through ad placement on the website, and provides two sub-fora: the experts' council and the user forum. It is only available in German. In the experts' council, users can ask questions and get responses from medical experts, depending on the thread in which they choose to post their message. In the user forum, which is the forum we selected for the project, people discuss their experiences with specific diseases or medication and help each other in specific life situations. Threads users post in are, for instance, *allergies*, *primary care*, or *infections*. The forum is moderated to keep conversations civil, and moderators sometimes point patients to specific threads in the experts' council subforum to get more detailed or short-termed information. We built a crawler and downloaded all posts available in the user forum in July 2021, containing posts between 2000 and 2021. All messages were filtered for Covid-19-related posts to remove potential vaccine-related reactions and discussions and avoid biasing our dataset towards this topic. We arrived at approximately 69,700 posts from Lifeline without any further pre-processing.

French

For French, we found it very difficult to receive access to any "real" data. For every potential resource, Requirement 3 would have been hurt. Thus, we resorted to translations of one part of the German data (for now). We translated a part of the data which was already de-identified

[8] https://chiebukuro.yahoo.co.jp/
[9] https://fragen.lifeline.de/forum/
[10] https://funkedigital.de/wer-wir-sind/

and annotated with binary labels (see Section 4.1.2), to reduce the annotation and curation effort.

We used DeepL[11], an online neural machine translation service, for translating the German texts into both English and French. Then, we provided the texts in all three languages to our French team members and to two French-speaking annotators[12]. The task of the French speakers was to check if the French texts were usable for our purposes without looking at the German original or English translation. We defined "usable" as follows:

1. The text is readable and understandable without reading the English or German version,[13] i.e., the reader can understand without effort what the patient intends to say.
2. The text does not need not be perfectly grammatically correct (the original documents aren't either).
3. The text "sounds" like it was written by a patient in an online forum.[14]
4. The text does not contain more than one word/phrase that does not make sense in this context, i.e., it only has minor issues that do not impede comprehensibility.

If the texts were mostly well translated, according to the above-described criteria, they would be marked as *checked*. As "minor issues," we regarded wrong or confusing translation chunks, which were most often based on typos or ambiguities in the original German document, as, for instance, can be observed in Example 4.1a and its (incorrect) translation in Example 4.2a.

(4.1) German original

a. ***de:*** *"... ich meine, dass* ***der Stuhl*** *nicht mehr so geformt ist ..."*
b. ***en:*** *"... I mean that the stool is no longer formed in such a way ..."*

(4.2) French (incorrect) translation

a. ***fr:*** *"... je veux dire que* ***la chaise*** *n'a plus la même forme ..."*
b. ***en:*** *"... I mean that the chair is no longer formed in such a way ..."*

This mistranslation in French is very clearly due to the ambiguity of the German word "Stuhl", which could mean both "chair" and "stool" (in the sense of feces) in English. In case of such occurrences, we asked our team members to flag these texts as *needs improvement* and *checked*. Texts that were utterly unintelligible, i.e., when significant improvements were needed, the texts were to be marked as *discard* and *checked*. Major improvements were defined as follows:

1. The text is not understandable when reading it for the first time.
2. The words in the original document seem to be translated literally throughout the text.[15]
3. The text contains more than one German / non-understandable phrase.

Of course, these definitions are rather vague. However, the goal was not to rate the quality of translations but to quickly find suitable French translations without developing and going through a checklist of criteria.

After adding these markers, our annotators were asked to go again through the examples flagged as "needs improvement" and improve the translations with minor flaws. For this, they

[11] DeepL allows free translation of 5,000 characters per month. URL: https://www.deepl.com/translator
[12] For more information about our annotators' backgrounds, please refer to Appendix B.2.
[13] Note that the English translations had a much worse quality than the French translations.
[14] We noticed when comparing social media posts in French, German, English and Japanese that they all "sound" very differently.
[15] This seemed to be often the case for the English translations.

could also check the German original version.[16] For instance, since Example 4.2a does not make much sense in the document context, it can be easily fixed, as shown in Example 4.3.

(4.3) ***fr:*** *"... je veux dire que* ***les selles*** *n'ont plus la même forme ..."*

Another example of an easy improvement was often observed for abbreviations, especially for greetings at the end of the user's post. For example, in some cases, the greetings at the end of a post were not correctly translated: "LG" in German is short for "Liebe Grüße" (***en:*** *"Kind regards / cheers / best"*, used in similar contexts as "bises" (*fr*, ***en:*** *"kisses"* in the literal translation) and more colloquial than "amicalement" (***en:*** *"amicably"* in the literal translation). Those could be simply fixed using a search-and-replace function. Other abbreviations required more effort for the translation, e.g., the acronym "HET" in the sentence "Ich habe vor ca 10 Wochen mit meiner HET angefangen" ("HET" is short for "Hormonersatztherapie", ***en:*** *"hormone replacement therapy (HRT)"*). DeepL did not translate the acronym, and therefore, our annotators took over the task, translating the sentence into "J'ai commencé mon traitement hormonal substitutif il y a environ 10 semaines" (***en:*** *"I started hormone replacement therapy about 10 weeks ago."*).

The annotators were asked first to check the *positive* texts, i.e., those that contained any ADRs. They then continued with the *negative* ones. This order was because we wanted to have relevant documents for entity and relation annotation as soon as possible. Currently, 864 documents have already been checked and, if necessary, manually improved. The annotation of the French documents then followed the procedure for the German documents.

De-Identification

We de-identified all documents using regular expressions[17]. Very common occurrences were, for example, user names. Often, users greet each other, "sign" their posts with their names, and refer to each other using nicknames (or nicknames of nicknames, for example "Mohnblümchen" for the user name "Mohnblume", a diminutive of "poppy"), thus they occur quite frequently. We collected the regular expression matches and replaced them with a mask (`<user>`) to keep the conversation structure intact. However, since users are very creative in inventing names, greetings, and goodbyes, not all of them were captured. Therefore, one of the tasks for the annotators was to add an entity label "user" to all still-existing names during the entity annotation process. Those were then replaced after annotation.

We proceeded similarly for URLs, (e-mail) addresses, and other personal information, to remove any trace of the post authors. However, since users sometimes name the city they live in or even their doctor's name, we cannot say for sure that each mention was captured. We, therefore, ask the annotators to mark any remaining PHI and replace it accordingly with the masks `<URL>` or `<pi>` (personal information). We also replace occurrences of exact dates or years with `<date>`.

4.1.2 Binary Annotation

First, the binary annotation process is described. The goal is to categorize documents as shown below into the labels *positive* (Example 4.4) and *negative* (Example 4.5). The binary annotation is only applied to the German corpus since we needed it to find the relevant documents for the more detailed annotations described further below. The German data is then translated into French, and the labels are taken over. For Japanese, a binary annotation was not necessary since the way the data were collected already reduced the number of documents to annotate.

(4.4) Example of a document labeled as *positive*:

[16] Both annotators speak German and English as well as French.
[17] This was done before the translation into French.

a. ***de:*** *"Hallo <user>, ich habe mein Sulfasalazin nach 4 Monaten erfolgloser Einnahme wegen starker Bauchschmerzen abgesetzt. Entzugsentscheinungen habe ich nicht bekommen. Liebe Grüße vom <user>"*
b. ***en:*** *"Hello <user>, I stopped my sulfasalazine after 4 months of unsuccessful use due to severe abdominal pain. I did not experience any withdrawal symptoms. Kind regards from <user>"*

(4.5) Example of a document labeled as *negative*:

a. ***de:*** *"Hallo <user>, das Mittel welches Du genommen hattest, war waohl der Vorgänger von dem Calmvalera. Ich habe Ängste und auch dadurch wohl diese innere Unruhe. LG <user>"*
b. ***en:*** *"Hello <user>, the medication you took was probably the predecessor of Calmvalera. I have anxiety and also probably because of this inner restlessness. Cheers <user>"*

Guideline Development

We decided first to annotate the posts on the document level to find those relevant to ADRs. The guidelines for those are straight-forward:

Definition 3 (Binary Annotation)**.** *A document is labeled as "contains one or several ADRs" (positive) if an adverse reaction is clearly stated and experienced by the user themself. In contrast, posts that do not describe any ADRs or only repeat other users' side effects are to be labeled as negative. Further, documents containing negated ADRs are negative as well.*

More ambiguous cases are, for example, documents where ADRs are only referred to or mentioned using expressions like "side effects" or only implicitly, like "could not tolerate". These are to be annotated as *positive*, too. Posts in which the users describe how close relatives, e.g., spouses or children, experience side effects, are *positive* examples as well. Documents where it is unclear which message the user wanted to convey are to be annotated as *negative*. Rumors and other speculations are to be labeled as *negative*. These guidelines can theoretically be applied to any language.

Annotation Process

We chose PRODIGY[18] for the binary annotation task because it is easy to set up and user (annotator) friendly. The binary labeling task can be displayed to the annotators as a simple clicking task: "accept" (*positive* document) versus "reject" (*negative* document), see an example in Figure B.1. This makes labeling very quick and intuitive. We configured the annotation tool to have a `feed overlap`, meaning that every annotator sees (and annotates) every document. We decided on this procedure because we found that even the binary annotation was sometimes more complex than anticipated.

Initially, only one annotator was available (Annotator 0).[19] He was trained with 200 English examples, which were discussed afterward. He then started on the German data. In weekly meetings, problems or ambiguities were discussed. As long as we did not have a second annotator, the author of this thesis also took part in annotating parallel examples.

Although the task of categorizing documents into the classes *positive* and *negative* does not seem very complicated at first glance, it turned out that some examples were very ambiguous. This was especially true for posts related to menopause where people describe symptoms that might be ADRs as well but are actually "normal" accompanying symptoms of menopause. See Figure 4.2 for an overview of topics discussed in the data; Documents containing discussions

[18] https://prodi.gy/
[19] More details on the annotators are provided in Appendix B.2.

related to women's health are strongly represented. Some other posts were merely repeating rumors, summarizing other users' reports to give them advice, or collecting information from the internet. The annotators also encountered documents in which users reported an ADR which happened to them in the past, leading the patients to change the medication with which they are now happy, that is, at the time of writing the text, there is no ADR present anymore. Many examples also appear to be speculative.

Therefore, in the beginning, much discussion was dedicated to these borderline cases. In the end, we resorted to always keeping the patient's perspective in mind, following the rule to mark documents as *positive* if *the person writing them* believed to have experienced an ADR, even if we did not think so. This is because none of us were experienced medical doctors, and catching more ADRs in the IE process would be better than missing any. Further, repetitions and information from somewhere other than the patient's own experience were labeled as *negative*. Speculations and the spread of untrue information related to health were flagged for later investigation.

After about 2.5 months following the binary annotation's start, we hired another annotator, Annotator 1. He was trained similarly to Annotator 1 and started annotation on the parallel dataset soon after. After Annotator 0 could not work with us any longer, we were able to employ a third annotator, Annotator 2. We repeated the training process as described above, and she took over the annotation started by Annotator 0.[20] In weekly meetings, we continued to discuss any emerging difficulties.

Curation was done with the PRODIGY `review` module, where all concurrent examples were automatically merged while both annotators and the annotation instructor reviewed the non-concurring ones. The final label was a majority vote on the opinions.

Note that sometimes, due to some inner workings of the annotation tool[21], some documents were only annotated by one of the annotators. If these showed up during curation, they were treated as a regular document and discussed to find a final label. The same was applied to documents that were flagged or ignored by one of the annotators, which sometimes happened when the annotators did not understand the texts or were unsure how to label them. We stopped binary annotations after about 10,000 documents were merged and consolidated.[22]

In Figure B.2, the annotation progress for the binary annotations is shown. It took between 8 and 9 months in total to annotate and curate approximately 10,000 documents[23]. After binary annotation and some pre-processing, e.g., removing documents with a length shorter than four tokens, we arrived at 10,010 annotated and curated documents. Of those, 9,689 documents were labeled as *negative* and 324 as *positive* as shown in Figure 4.1. Thus, there are only 3.2% of the documents in our dataset contain ADRs, similar to related work.

Inter-Annotator Agreement

We calculated the IAA comparing annotators' labels with our final gold standard and against each other. In the binary case, we can count the number of negative samples and therefore, we can use the κ statistic as a metric. See Table 4.1 for the achieved scores with respect to the discussed IAA metrics (Section 2.2.2 provides more details on the respective scores). Unfortunately, they show very low agreement for Cohen's κ (0.19). The macro average F_1 score is in

[20] From here on, we assume for the sake of clarity that Annotator 0 and Annotator 2 are the same person, even when calculating IAA.

[21] Most probably a bug in PRODIGY in how documents are presented to the annotators, which was not detectable before curation.

[22] Note that we published some first results, using approximately half of the binary annotated data, in **Raithel et al.** (2022). However, we kept annotating the data until we reached 10,000 documents, which will be released in a follow-up paper.

[23] The annotators worked for about 10 hours a week.

the middle range and observed agreement seems quite good with a score higher than 0.90. Table 4.1 further demonstrates the variety of resulting IAA scores when considered from different perspectives.

metric	score
Cohen's κ	0.19
F_1 score (macro avg)	0.59
observed agreement	0.96

Table 4.1: The results for the different agreement scores of binary annotation when comparing the annotators with each other. F_1 score is shown as macro average over both classes.

The confusion matrices in Figure 4.1 show the number of examples both annotators agreed upon (top left) as well as a comparison of each annotator with the gold standard data (top right and bottom left) and the final number of documents per label. Note that some of the documents were only annotated by one annotator; therefore, the numbers do not add up to 10,013 for all matrices. Annotator 1 and Annotator 2 agreed on the *negative* label of 9,057 documents, while they only agreed on 47 documents to be *positive*. Furthermore, Annotator 1 labeled 111 documents as *positive*, while Annotator 2 rejected these. On the other side, Annotator 1 rejected 255 documents that were accepted by Annotator 2.

Figure 4.1: Confusion matrix for the binary annotation of 10,013 documents in total. Upper left: annotator 1 (a1) versus annotator 2 (a2) annotation decisions. Upper right: a1's annotation decision versus the final gold standard. Bottom left: a2's annotation decisions versus the final gold standard. Bottom right: The final gold dataset with 9,689 *negative* and 324 *positive* documents. "reject" translates to *negative*, while "accept" translates to *positive*.

In the top right and bottom left matrices in Figure 4.1 it can be observed that Annotator 1 agreed on the label of 9,147 *negative* and 96 *positive* documents with the gold standard. These numbers are significantly higher for Annotator 2, who agreed on 9,594 *negative* and 256 *positive* documents. Annotator 1 agreed with the gold standard on 9,244 annotations, while Annotator

2 agreed on 9,851 annotations. The observed agreement *with the gold standard* and without correction for chance is therefore 0.92 for Annotator 1 and 0.98 for Annotator 2. Comparing the annotators with each other resulted in a observed agreement of 0.96. Finally, we also calculate classification scores as used in the evaluation of classification systems, to show the agreement per class. These are shown in Table 4.2.

		precision	**recall**	**F_1**	**support**
	negative	0.98	0.99	0.99	9178
	positive	0.57	0.31	**0.40**	292
a1 vs. gold	accuracy			0.97	9,470
	macro avg	0.77	0.65	0.69	9,470
	weighted avg	0.97	0.97	0.97	9,470
	negative	1.00	0.99	0.99	9178
	positive	0.82	0.85	**0.84**	292
a2 vs. gold	accuracy			0.99	9,470
	macro avg	0.91	0.92	0.92	9,470
	weighted avg	0.99	0.99	0.99	9,470
	negative	0.97	0.99	0.98	9168
	positive	0.30	0.16	**0.20**	302
a1 vs. a2	accuracy			0.96	9,470
	macro avg	0.64	0.57	0.59	9,470
	weighted avg	0.95	0.96	0.96	9,470

Table 4.2: Inter-annotator agreement when using F_1 score score, i.e., comparing each annotator's (a1 and a2) labeling with the final gold label and comparing the annotators with each other (bottom). We used only those samples for calculation for the two upper rows for which both annotators assigned "accept" (*positive*) or "reject" (*negative*). The bottom row shows the agreement of annotators with each other, taking the annotations of Annotator 1 as gold labels. Here, the number of accepted documents is higher than in the two upper rows. We highlight the scores for the positive documents since these are the most relevant ones for our work.

Analysis If we take a closer look at the agreement of the *positive* documents, we see that Annotator 1 clearly agreed on fewer *positive* documents with the gold standard (96 in total) than Annotator 2 (256 documents). This shows the difficulty of annotating the documents with only a binary label.

Regarding Cohen's κ, the score used in other works describing the creation of a corpus containing documents with ADRs, we find that all of them exceed our IAA score ($\kappa_{Cohen} = 0.19$): Klein et al. (2020) achieved a κ_{Cohen} of 0.49 for Russian Twitter messages, and a κ_{Cohen} of 0.61 and 0.69 when comparing three annotators of French Twitter messages, and Zolnoori et al. (2019) report $\kappa_{Cohen} = 0.78$ for English forum posts. In contrast, the scores we calculated using, for instance,F_1 score are decently high when calculated on the *entire* corpus ($F_{1(weighted)} = 0.96$). This has two reasons: First of all, F_1 score does not consider chance agreement, as Cohen's κ does, and second, the vast data imbalance basically eradicates the contribution of the scores on the positive documents. When looking at the agreement between annotators on only the positive documents, we also see a very low score of $F_{1(positive)} = 0.20$. The macro average F_1 score of 0.59 takes the label imbalance into account (but not agreement by chance), and is still relatively low.

These low agreement scores demonstrate that it is very easy to guess a label for a document and be correct randomly: Most documents are negative, so a random guess of the negative label does not hurt much. Since Cohen's κ in Table 4.1 accounts for randomness, they result in a very low score. Intuitively, however, it is very hard to guess a positive label and be correct with it randomly.

This shows that neither of the presented scores can accurately measure the agreement on an imbalanced corpus such as the one presented. Also, in contrast to many other works on ADRs, we created a dataset of approximately 10,000 documents, where (almost) every document was annotated by two annotators. We further curated all documents where no agreement was reached.

Final Dataset annotated with Binary Labels

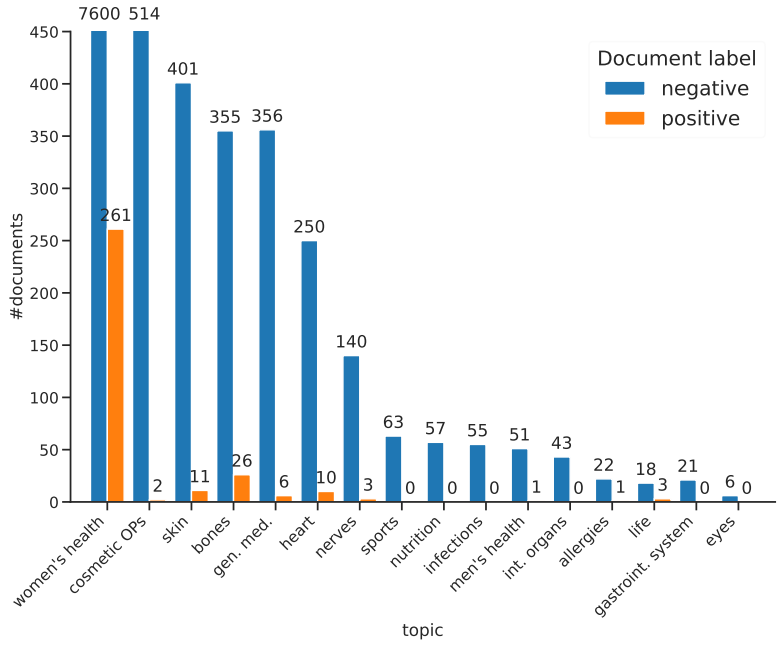


Figure 4.2: The topic distribution of the annotated LIFELINE-DE-ALL documents (de). For a better visualization, the y-axis is cut off at 450 documents.

The final datasets with binary annotation are summarized in Table 4.3. In total, the LIFELINE-DE-ALL corpus comprises 10,013 documents, of which 324 documents are positive. This corpus is divided into LIFELINE-DE-1 and LIFELINE-DE-2. LIFELINE-DE-1 was translated into French and also used as the dataset for the experiments in Chapter 5. Translations that were not understandable after translation were removed. Therefore, the new French dataset with binary annotations consists of 864 documents, with precisely 100 documents being positive.

German We split the single posts into sentences and tokens using spacy[24] to gain some insights into the dataset. Stop words were filtered out and all tokens were lower-cased, but not stemmed[25]. The dataset contains approximately 47,600 unique tokens (not lemmatized). Some of the most used words, apart from "hello" and similar expressions used in online fora, are "Angst" (***en:*** *"anxiety, unrest, worry, apprehension"* depending on context, 1,115 occurrences), "Arzt" (***en:*** *"doctor"*, 1,107), "Beschwerden" (***en:*** *"afflictions, discomforts"*, 966), "wünsche" (***en:***

[24] https://spacy.io/

[25] Note that the lowercase might merge some tokens that do not have the same meaning.

name	language	#documents	#positive	#negative
LIFELINE-DE-1	de	4,169	101	4,068
LIFELINE-DE-2	de	5,844	223	5,621
LIFELINE-DE-ALL	de	10,013	324	9,689
LIFELINE-1-FR	fr	864	100	764

Table 4.3: Overview of the binary annotated data. LIFELINE-DE-ALL is the combination of LIFELINE-DE-1 and LIFELINE-DE-2. LIFELINE-1-FR is the current state of translations of LIFELINE-DE-1, where unintelligible translations were removed, thus the lower number of documents in total.

"wish", 869), "WJ" (an abbreviation of "Wechseljahre", ***en:*** *"menopause"*, 819), and "Hormone" (***en:*** *"hormones"*, 767). This reflects a typical story told by many patients in this forum: Many people are scared of what might happen if or if they do not take a drug, afraid of receiving negative results of medical tests or worrying about a person close to them. Further, many go from one physician to the next without getting any better. Often, this concerns women in their menopause, which is why "menopause" and "hormones" are also mentioned frequently. Indeed, this is the most discussed topic, as can be seen in Figure 4.2, where we count 7,600 negative and 261 positive documents, by far the highest number of documents for both labels. Note that the word "Nebenwirkungen" (***en:*** *"side effects"*) and its variations occur 390 times. Interestingly, only 96 occurrences are allotted to the positive documents, the rest (294 mentions) occur in the negative documents. These might contain some negated constructions.

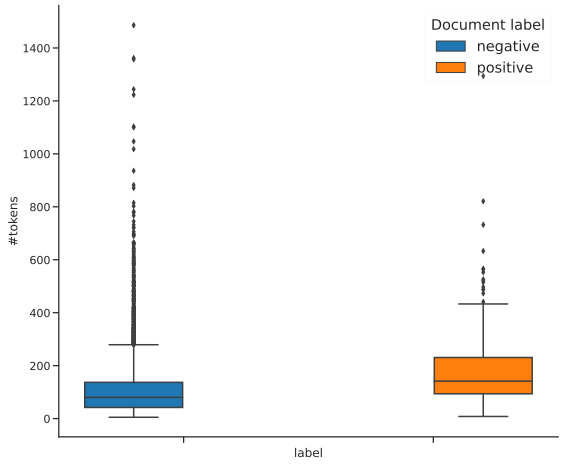


Figure 4.3: The distribution of the number of tokens per document and per label in the LIFELINE-DE-ALL corpus (de). For a better visualization, we removed the longest document containing 3,166 tokens.

An often-seen phenomenon in corpora of documents containing ADRs is that positive documents are longer than negative ones. This intuitively makes sense, since people who describe their problems tend to write more than when everything is alright and they just forward some information to other patients. In both Figure 4.3 and Figure B.3, we can see a slight tendency towards exactly that, for both the number of tokens and the number of sentences, respectively. However, there are also some very long negative documents. This might be due to the forum being more of a general platform rather than Twitter (people not experiencing side effects are maybe less likely to post about their medication intake) or drug review forums, which are mostly geared towards detailed reports containing negative experiences with respect to drugs.

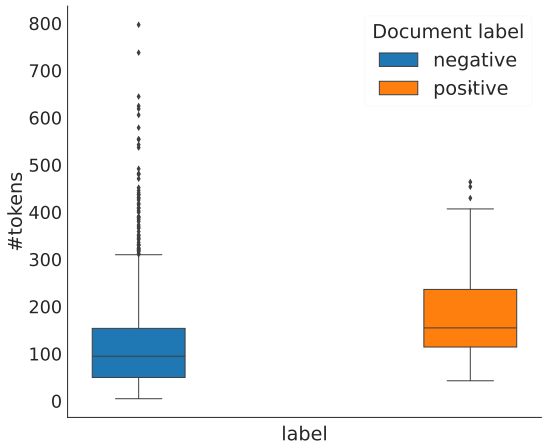


Figure 4.4: The distribution of the number of tokens per document and per label in the LIFELINE-1-FR corpus (fr).

French The French corpus is much smaller and shows a different label distribution since we ensured that most positive documents were translated while unintelligible translations of the negative documents were discarded. Also, since the project is ongoing, not all translations have been reviewed yet. Table 4.3 shows the current state of the data at the time of writing: We currently have 100 *positive* and 864 *negative* documents that were translated and improved if necessary. These are used for the following, more detailed annotations. The distribution of tokens per document across labels is shown in Figure 4.4.

4.1.3 Entity and Relation Annotation

The guidelines for annotating entities and relations are more involved since they are not based on a yes-no question, which can be asked and answered independently of the language. First, they need more fine-grained distinctions, e.g., to find the exact span in a text that describes the ADR or an opinion about a medication. For this, deciding which parts of a document are relevant to our goal is also necessary. Further, the guidelines need to be applicable to at least three languages (plus English) and capture all relevant constructions occurring in these languages. And finally, all project partners had different experiences and foci in mind which had to be united.

The guidelines were developed by first using random English examples from the CADEC (Karimi et al., 2015) corpus annotated by all KEEPHA teams to find the first potential entities of interest. The annotate-then-discuss circle was repeated several times, while each team also tried to find (or make up) corresponding or complex/exceptional cases in their own language to test for various expressions. In addition, we used translations of the CADEC documents as annotation dummies. Note that these were often much simpler than the ones we later encountered with our own data. This thesis briefly outlines the major points and difficulties in developing the guidelines. The document containing the complete annotation guidelines is available online[26].

See an example annotation below, which already shows some differences between the original German version (Example 4.6) and the English translation (Example 4.7): While in German, "Haarausfall" (***en:*** *"hair loss"*) is a compound, it consists of two nouns in English. Also, in the English version, we can annotate only the medication mention ("MTX"), but for German, we

[26] https://cloud.dfki.de/owncloud/index.php/s/NdrN9y9EJQ39b77

have to take the entire span ("MTX-Einnahme", ***en:*** *"MTX intake"*), since embedded entities are not allowed, according to our guidelines.

(4.6) ***de:*** ... die Erfahrung mit Haarausfall$_{\texttt{disorder}}$ bei MTX-Einahme$_{\texttt{drug}}$ erlebe ich .. (caused)

(4.7) ***en:*** ...also experiencing hair loss$_{\texttt{disorder}}$ with MTX$_{\texttt{drug}}$ intake .... (caused)

Annotation Guidelines: Entities

The set of entities for the annotation scheme was developed step by step. The final set of entities is shown in Table 4.4. When one of the teams (i.e., the annotators and annotation instructors) found that the provided entities did not capture a relevant medical concept in the English and translated test examples, the examples were discussed in the next KEEPHA meeting. If the concept seemed to be relevant to the other languages as well, it was investigated with further examples and, if judged necessary by a majority, incorporated into the annotation scheme and guidelines. In some cases, it was difficult to transfer the meaning of special cases from one of the languages to English and convey the meaning to the project partners.

In general, entities are annotated by marking the noun or verb phrases, if present with their modifying parts, e.g., adjectives that describe the entity of interest in more detail. The annotators were instructed to annotate the smallest phrase possible. However, since some descriptions are rather long, a common characteristic of user-generated text, we also allowed longer phrases if necessary. Moreover, metaphoric expressions were to be labeled as well, since these also account for a big part of the expressions people use to describe their ailments. Determiners and possessive pronouns were not annotated, and enumerations were to be split up and annotated separately.

We did not allow nested entities, to make automatic processing easier and avoid confusion. However, how to handle this mainly depends on the language. For instance, the German word *"Kopfschmerzen"* (***en:*** *"headache"*) is a compound (similar to the English translation) and should therefore not be split up in *"Kopf"* (***en:*** *"head"*) and *"Schmerzen"* (***en:*** *"pain"*). For the French *"douleur thoracique"* (***en:*** *"chest pain"*) we proceed exactly the same, although more complex constructions, like *"douleur au thorax"* (***en:*** *"chest pain"*) are annotated separately: *"douleur"* (***en:*** *"pain"*) is marked as a `disorder`, while *"thorax"* (***en:*** *"chest/thorax"*) is labeled as `anatomy`. In contrast to nested entities, we *did* allow discontinuous entities in case there is no other way to catch a specific expression. An example is shown in Example 4.8 and its English translation in Example 4.9.

(4.8) ***de:*** 3kilo$_{\texttt{disorder}}$ habe ich auch seitdem runter$_{\texttt{disorder}}$ ... (discontinuous)

(4.9) ***en:*** I have also lost$_{\texttt{disorder}}$ 3 kilos$_{\texttt{disorder}}$ since then ... (discontinuous)

In the English translation, having a discontinuous entity is unnecessary, since "lost 3 kilos" can be annotated as a whole. However, the syntax is different in the German original, and the relevant entity parts are split apart.

As in related work, we included spelling mistakes and colloquial language in the annotation but excluded punctuation markers. We did not pre-process the data to correct any noise or errors. Abbreviations were annotated since they represent another common occurrence in

the underlying data. In addition, we also assigned attributes to provide more fine-grained information on the relevant entities. They will be explained with their respective entity labels.

Annotation Scheme: Entities

entity type	attributes
drug	increase, decrease, stopped, started, unique_dose
change_trigger	
disorder	negated
function	negated
anatomy	
test	
opinion	positive, negative, neutral
measure	
time	frequency, duration, date, point
route	
doctor	
user	
url	
personal_info	
other	

Table 4.4: Overview of the different entity types and their attributes. Note that user, url, and personal_info are only annotated for de-identification purposes.

As for the entities themselves, the most important are drug, disorder and function. Drug represents any mention of a medication name (abbreviated or not), a brand, or an agent. Also, dietary supplements are included, as well as drug-based treatments and mentions referring more generally to drugs, for instance *"Arthritis medicines"*. Drug entities might have one out of five attributes: increase, decrease, stopped, started, unique_dose (see Example 4.11). These describe if the medication was, for instance, just started, or if the dosage was increased.

Another entity called change_trigger is related to that: Certain expressions in the text might lead to a change in the status of medication intake, for instance "to start" or "to taper off". These are labeled as well, to give more information on the current status of the medication.

Disorder is the label for any sign, symptom, or disease expressed by the patients. These can be very long and are often expressed via metaphors or implicitly. Very broad and non-specific phrases, like *"I do not feel well"*, are also marked as disorder. While disorder usually describes a malfunction of the patient's body, function is the label for neutral or positive processes of the body, including mental functions. Consequently, a negated function is a disorder and should be annotated as such. We provide more detailed guidance on how to distinguish between disorder and function in the guidelines. Example 4.10 and Example 4.11 also show two text snippets containing these and more entity types.

(4.10) Disorder:

a. ***fr:*** *"J'ai une* *maladie de crohn*$_{\texttt{disorder}}$ *depuis 36 ans*$^{\text{duration}}_{\texttt{time}}$ *(...)"*

b. ***en:*** *"I've had* *crohn's disease*$_{\texttt{disorder}}$ *for 36 years*$^{\text{duration}}_{\texttt{time}}$ *(...)"*

(4.11) Function:

a. ***de:*** *" Opripramol*$^{\text{stopped}}_{\texttt{drug}}$ *hatte*$_{\texttt{change\_trigger}}$ *ich ja nur zwei Abende lang*$^{\text{duration}}_{\texttt{time}}$ *zum Schlafen*$_{\texttt{function}}$ *je eine halbe*$_{\texttt{measure}}$ *genommen, also nur eine winzige Dosis*$_{\texttt{measure}}$ *."*

b. ***en:*** *"I had*$_{\texttt{change\_trigger}}$ *only taken half a dose*$_{\texttt{measure}}$ *of Opripramol*$^{\text{stopped}}_{\texttt{drug}}$ *for two evenings*$^{\text{duration}}_{\texttt{time}}$ *or sleeping*$_{\texttt{function}}$ *, so only a tiny dose*$_{\texttt{measure}}$ *."*

Negation is an attribute that is (currently) only applicable to the labels `disorder` and `function` (`negated`), since these are the most interesting for our use case and it is important to know if a disorder does not exist anymore. Note that in contrast to existing corpora, we annotate all kinds of symptoms (disorders), not only those related to drugs or those that express an ADR.

The entity label `anatomy` refers to any part of the body, also, for example, hair and nails. Anatomical entities are not annotated separately, however, in case they are part of a larger entity, e.g., in *"headache"*. `Test` marks all medical tests or examinations that produce a result that is used in a medical diagnosis, for example, *"blood test"*. The entity label `opinion` provides a way to mark the *writer's* opinion or evaluation of a certain drug, health state, or biological process. It is used with the attributes `positive`, `negative`, and `neutral` to denote the sentiment of the opinion (see Example 4.12). Note that we annotate all opinions found in the texts: those of doctors as well as those of relatives. We found this to be easier for the annotators. The patient's opinion is then further expressed by relations. `Measure` is an entity label that is used for all occurrences of drug dosages or test results, anything that is clinically relevant. A complex entity label is `time`. We use it for mentions of `frequencies`, `durations`, `relative points in time` or `dates` (see Example 4.10). Another entity label to give more fine-grained information on medication intake is `route`. This label is supposed to be annotated for all means by which a drug can be consumed, e.g., by injection, or by using pills. Often, these means are described using verbal phrases, but also nouns. The label `doctor` is used to add a marker for profession names, e.g., *"cardiologist"*. For any other entity that might seem relevant but where we did not define a concrete label, the annotators are instructed to use the label `other`. These expressions are to be investigated after annotation.

Finally, we added some entity labels that are used to further de-identify the corpus. This concerns mentions of user names (`user`), URLs (`url`) and any other personal information, e.g., addresses, city names, doctors' names etc. After the corpus is consolidated, these are replaced by corresponding masks, and the labels are removed.

Annotation Scheme & Guidelines: Relations

After having developed the final set of entities, we added the relations that associate these entities. In general, we do not annotate relations if none of the mentioned entities is concerned, if the document itself is hypothetical or speculative or if the relevant part of the document is formulated as a question.

The relation which connects the most important entities for our use case is `caused`. It associates a head argument, either `drug`, `function` or `disorder` with a tail argument, either `disorder` or `function`. In the guidelines provided above we define all these relations in more detail. Our approach stands in contrast to other works which often distinguish the symptoms/disorders based on their type (for instance, *indication* versus *adverse reaction*), while our annotation scheme describes ADRs only via the relation between a `drug` and `disorder` (or `function`) since symptoms and signs do not necessarily need to be ADRs.

relation type	head argument	tail argument
caused	drug, disorder	disorder, function
treatment_for	drug	disorder, function
has_dosage	drug	measure
experienced_in	disorder	anatomy
examined_with	disorder, anatomy, function	test
has_result	test	measure, disorder, function
refers_to	disorder	disorder, function^{negated}
refers_to	drug	drug
refers_to	anatomy	anatomy
refers_to	function	function
interacted_with	drug	drug
signals_change_of	change_trigger	drug
has_time	drug, disorder	time
has_route	drug	route
is_opinion_about	opinion	drug, disorder, function
misc	ANY	ANY
not tracked	URL, personal_info, user	

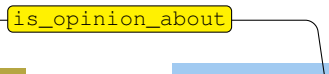
Table 4.5: Overview of available relation types and the entity types they associate.

The opposite of that is the relation called treatment_for, which, again, connects a medication with a disorder or function. This time, however, the medication is meant to treat the disorder in some way. The relation has_dosage connects a drug with a measure, so describe the dosage of the medicine, if existent. If a symptom is felt in a certain part of the body, the mention of the disorder is connected to the anatomy entity using the relation experienced_in. To connect the test entity with the entity that was tested (disorder, anatomy or function), we introduce the relation examined_with. The result of this test might be mentioned as well and can be associated with the test's result using the relation label has_result. The test's result might be a measure, disorder, or function.

The relation refers_to is a means to reduce the number of relations to the same head or tail argument. Several mentions of the same entity, e.g., a drug can be connected via this relation and only one of them needs to be associated with the other entities of interest, e.g., a disorder. To mark drug-drug interactions, we introduce the label interacted_with. These can often be the reason for a disorder. Further, for connecting the change_trigger with the drug, we use the relation signals_change_of. Drug and disorder expressions can be connected with has_time to time markers to express, for example, a frequency of drug intake. The same is valid for the relation has_route, which connects a drug with its route. If a patient states a specific opinion with respect to a drug, disorder or function, it can be related to those entities with the label is_opinion_about, as shown in Example 4.12[27] and its translation in Example 4.13.

(4.12) ***fr:*** ... à prendre la mirtazapine$^{started}_{drug}$, (...), et je me suis sentie beaucoup mieux$^{positive}_{opinion}$. (is_opinion_about)

[27] This sentence is shortened to allow for better visualization. Originally, it reads ***fr:*** *"J'ai pris un antidépresseur pendant un certain temps, la mirtazapine - la plus petite dose. L'année dernière, à Noël, j'ai recommencé à en prendre, même un demi-comprimé seulement, et je me suis sentie beaucoup mieux."*, ***en:*** *"I took an antidepressant for a while, mirtazapine - the lowest dose. Last year, at Christmas, I started taking it again, even just half a tablet, and felt much better."*

(4.13) *en:* ... to take mirtazapine$^{started}_{\texttt{drug}}$, (...), and felt much better$^{positive}_{\texttt{opinion}}$. 

Finally, for everything else that seems relevant but for which no relation exists so far, we can use the relation type `misc`.

Annotation Process

For annotating the German part of the KEEPHA dataset with entities and relations, we first selected all *positive* documents within the LIFELINE-DE-2 corpus (see again Table 4.3) and divided it into two batches of 118 and 105 documents. The annotators started with the batch containing 118 documents.

Since we also want to annotate entities and relations in the French data, which is based on LIFELINE-DE-1, we want to prevent dataset leakage by annotating the same documents, even if they are translated. Therefore, the French data for entity and relation annotation was taken from the translated corpus (Lifeline-v1-fr, see Table 4.3), resulting in 100 positive documents. We randomly sampled some negative examples as well for a later annotation.

Using PRODIGY, we quickly annotated some (simple) `disorders` and `drug` occurrences using a basic German model with active learning in the background. This so-trained model was then applied to the data. These data, including the pre-annotations, were then converted to BRAT format[28]and given to the annotators with the aim of reducing the annotation burden, especially for repetitive and frequent mentions.

The same annotators that already worked on the binary annotation were again trained on English data and on some German samples. The German data was then provided to the two German native speakers, and the French data was provided to our French-speaking annotators.

Each pair of annotators was provided with the same documents. We gave them the choice of either annotating first entities and then relations or annotating both at the same time. After completing the first German batch, we consolidated the data using a script provided in the BRAT library. It merges agreeing annotations and highlights disagreements. These were resolved in a final round of annotations. An overview of the data as of the time of writing is given in Table 4.6.

name	#documents	#entities	#relations	#attributes	comment
KEEPHA-de-part1	118	3,487	2,163	1,141	curated
KEEPHA-de-part2	105	-	-	-	in progress
KEEPHA-fr-1	100	1,939	1,129	537	A1 done

Table 4.6: Overview of the data annotated with entities, relations, and attributes. For the French data, only annotator 1 (A1) has finished annotation, therefore, the statistics refer to a non-curated version of the dataset. Note that this dataset will grow in the future.

Inter-Annotator Agreement

IAA was calculated for all annotation levels, that is, entities, attributes, and relations. Note that for each level, we calculated a strict and a relaxed version. The strict measure accounts only for annotations where the entity span matches between annotators, the relaxed measure also accepts overlapping spans, that is, spans that do not match perfectly. However, since all three

[28]We agreed on using BRAT in the KEEPHA consortium since everyone was familiar with it and it was easy to set up. Further, it allows to annotate attributes.

levels were annotated at the same time and only curated afterwards, errors made in the first level (entities) are propagated to the next levels.

At the time of writing, the French annotations were only finished by one annotator, so we do not calculate agreement for these annotations, but only for the first batch of German data. We provide the relaxed scores of entity, relation, and attribute annotation agreement in Table 4.7, Table 4.8 and Table 4.9. The strict scores, together with the counts of TPs, FNs, and FPs are presented in Appendix B.3.

	relaxed		
entity type	**Precision**	**Recall**	**F_1**
`anatomy`	0.49	0.66	0.56
`change_trigger`	0.53	0.51	0.52
`disorder`	0.80	0.87	**0.84**
`doctor`	0.79	0.94	**0.86**
`drug`	0.93	0.93	**0.93**
`function`	0.51	0.69	0.59
`measure`	0.66	0.78	0.71
`opinion`	0.15	0.53	0.23
`other`	0.24	0.38	0.29
`route`	0.56	0.47	0.51
`test`	0.55	0.59	0.57
`time`	0.85	0.55	0.67
micro average	0.75	0.79	0.77

Table 4.7: The relaxed IAA for entity annotation in the German data with the micro average scores across all entities in the bottom line. The three best F_1 scores are bold-faced.

The agreement in micro average F_1 score of the entity annotation is 0.77. The annotation of drug mentions is the one with the highest agreement ($F_1 = 0.93$), followed by mentions of doctor's professions ($F_1 = 0.86$) and disorders ($F_1 = 0.84$), which are rather good scores. The lowest agreement was found for the `opinion` and `other` entity types (0.23 and 0.29).

As mentioned before, we calculate the IAA for relations and attributes not on gold entities, but directly on the "raw" annotated data. With respect to the IAA on relation annotation, Table 4.8 shows a high fluctuation between relation types but also within relation types when taking a closer look at the head and tail arguments. The micro average F_1 score is 0.38. The highest agreement can be found for the `caused` relation when associating a `drug` with a `disorder` ($F_1 = 0.60$), which is our representation of ADRs. The two next best scores fall on the `treatment_for` relation between a `drug` and a `disorder` ($F_1 = 0.41$) and the `caused` relation connecting a `drug` and a `function` mention ($F_1 = 0.39$). For some relations, however, we do not find any agreement at all, e.g., `has_result` or `interacted_with`. All other agreement scores are rather low.

The IAA regarding attributes results in a micro average of 0.41 across all attributes. For negation and time attributes, Table 4.9 shows an F_1 score of 0.53 and 0.47 and almost no agreement on drug and opinion attributes.

Finally, we calculated the agreement between each annotator and the gold standard, that is, the final curated data. We report the relaxed scores. For entity annotation, the micro F_1 score is 0.80 for each annotator. This is 3 percentage points more than when comparing the annotators with each other. For relation annotation, the comparison with the curated data results in 0.47 and 0.45, respectively, which is an increase of 9 and 7 percentage points. Lastly, the scores for attribute annotation are 0.59 and 0.46 for each annotator, an increase of 18 and 5 percentage

			relaxed		
relation type	**head argument**	**tail argument**	**Precision**	**Recall**	**F_1**
caused	disorder	disorder	0.14	0.29	0.19
caused	disorder	function	0.00	0.00	0.00
caused	drug	disorder	0.62	0.57	**0.60**
caused	drug	function	0.00	0.00	0.00
caused	function	disorder	0.62	0.29	**0.39**
caused	function	function	0.50	0.25	0.33
experienced_in	disorder	anatomy	0.38	0.35	0.36
has_dosage	drug	measure	0.50	0.03	0.05
has_result	test	disorder	0.00	0.00	0.00
has_result	test	function	0.50	0.11	0.18
has_route	drug	route	0.25	0.04	0.07
has_time	disorder	time	0.75	0.09	0.16
has_time	drug	time	0.54	0.10	0.16
interacted_with	drug	drug	0.00	0.00	0.00
is_opinion_about	opinion	disorder	0.00	0.00	0.00
is_opinion_about	opinion	drug	0.11	0.11	0.11
is_opinion_about	opinion	function	0.00	0.00	0.00
signals_change_of	change_trigger	drug	0.43	0.19	0.27
treatment_for	drug	disorder	0.51	0.35	**0.41**
treatment_for	drug	function	0.00	0.00	0.00
micro average			0.47	0.31	0.38

Table 4.8: The relaxed IAA for relation annotation. The best three scores are highlighted and the micro average scores are shown in the bottom line. Note that (i) MISC and (ii) REFERS_TO relations were removed, since these are (i) for later investigation and do not have specific rules and (ii) subjective

points, respectively. According to these scores, annotator 1 was closer to the gold standard data than annotator 2, especially for the annotation of attributes. The increase in scores when compared to the curated data also confirms a phenomenon we observed during curation: The annotators are not annotating "wrong", i.e., giving wrong labels to the mentions, but they are not annotating thoroughly enough. The curation often resulted in a combination of both annotators' annotations.

Entities In the related work mentioned in Section 3.4, only Segura-Bedmar et al. (2014) use F_1 score for calculating IAA of their annotated entities. They report an F_1 score for the drug entity of 0.89 and 0.59 for ADRs mentions. Since we do not have an entity-specific to ADRs, we can only compare with the drug mention IAA, which is very close to ours ($F_1 = 0.93$) with only four percentage points difference. It is good that both disorder and drug receive a relatively high agreement, since these are our most crucial entities, together with the caused relation: These are the ones that represent ADRs.

When consolidating the annotations, we found several reasons for the low agreement scores on the remaining entity types. For instance, anatomy is sometimes hard to spot. Often, a disorder does not affect exclusively, e.g., the head of a person, but maybe a more specific region, like "the upper part of both legs". This, for instance, can lead to span disagreements. We also found that sometimes the annotators were not consistent in annotating all mentions of anatomy occurrences. Sometimes, a body part is not relevant to any medication or disorder, but should still be annotated for consistency, for example as in the sentence in Example 4.14

	relaxed		
attribute type	**Precision**	**Recall**	**F_1**
`negation`	0.68	0.43	0.53
`drug`	0.05	0.27	0.08
`opinion`	0.63	0.08	0.14
`time`	0.39	0.61	0.47
micro average	0.36	0.48	0.41

Table 4.9: The relaxed IAA for attribute annotation in the German data.

(4.14) a. ***de:*** *"Also nicht den* *Kopf*$_{\texttt{anatomy}}$ *hängen lassen immer motiviert an die Sache heran gehen."*
b. ***en:*** *"So don't hang your* *head*$_{\texttt{anatomy}}$ *(and) always approach the matter with motivation"* (literal translation)
c. ***en:*** *"So keep your* *chin*$_{\texttt{anatomy}}$ *up (and) keep yourself motivated."* (more natural translation)

`change_trigger` is another entity where especially the span is often difficult to pinpoint. Furthermore, potential triggers might be very implicit and therefore often missed by either annotator. However, on the other hand, we found during curation that if we are able to assign a change attribute to a drug mention, then there is almost always a trigger close by, so not annotating them can also come from simply forgetting about them. This might be due to the high number of entities and relations we have. In addition, the annotation interface can get slightly overwhelming and confusing as soon as some entities and relations are annotated. An example is shown in Figure B.5.

`function` is a difficult mention per se and the rather mediocre F_1 score is not entirely surprising. For example, "Wechseljahresbeschwerden" (***en:*** *"menopausal complaints, menopausal symptoms"*) can be both a `function` or a `disorder`, which is more clear in the English translation. "Symptoms" might be normal phenomena of menopause, but "complaints" rather point in the direction of having negative experiences, which are not "normal" anymore, and therefore a `disorder`. In general, the guidelines state that adverse or non-functioning biological processes are annotated as `disorder`, while neutral or positive processes are labeled as `function`. Further, negated functions are usually labeled as `disorder`, but this is not always the case, especially not for Japanese, which is why explicitly negated functions are allowed as well. Thus, the entity type depends strongly on the context and how the patients formulated their experience, which means that it is also related to the sentiment exuded by the text. However, this entity type might need a more strict definition and more training examples to make it easier for the annotators to decide between `disorder` and `function`.

In contrast, `measure` seems to be defined well enough with an F_1 score of 0.71, although we find a variety of expressions that can be seen as measurement. One example is shown below (Example 4.15). `test`, on the other hand, is difficult as well, since most patients do not describe exactly that they did some kind of medical test to investigate a specific symptom and then report the result, but they directly report the result, implying that they did a test, as in Example 4.15. Often, these implicit mentions were missed by the annotators.

(4.15) a. ***de:*** *"In dieser Zeit ist ja der* *P-Wert*$_{\texttt{test}}$ *generell niedrig, mein Wert war* *eher im unteren Bereich angesiedelt*$_{\texttt{measure}}$ *."*

b. ***en:*** *"During this time, the P-value$_{\texttt{test}}$ is generally low; my value was rather in the lower range$_{\texttt{measure}}$."*

The entity type `opinion` is a very subjective span to annotate. Some expressions can be interpreted as opinions, while others are rather general statements. However, when revising non-agreeing annotations, we found that many more spans could have been annotated as "personal assessments" of the patients with respect to a drug or their general health state and assumed that this is due to some slackness in annotation. During curation, we therefore tried to add more opinion expressions.

`other` is a category that collects all mentions that might be relevant to the medical "story" of the patient. Judging from the annotations, what is relevant is very subjective and differs a lot between annotators. Spans annotated by both of them include mentions of hospital stays, surgeries, therapies, and so on.

The type `route` is another difficult one. In some cases, it is not clear if a person is talking about the pharmaceutical form of a medication or simply uses the route name, e.g., "pills", to refer to a drug mentioned before. However, in most cases, the context determines if a drug is mentioned or if the patient really means the route. In Example 4.16, an example is shown that might be confusing at first. If we take the mention of "Tabletten" (***en:*** *"pills"*) literally, then the obvious entity type would be `route`, since pills are some kind of pharmaceutical form. However, the person actually means in this sentence that they are scared of *medication* in general (which often leads to not taking any), and not of pills as a route in particular.

(4.16) `drug` versus `route`

a. ***de:*** *"Hab größten Respekt vor Tabletten$_{\texttt{drug}}$ überhaupt, da ich weiß, was sie aus einem machen können."*

b. ***en:*** *"Have the greatest respect for pills$_{\texttt{drug}}$ in general, knowing what they can turn you into."*

The rather low agreement on the `time` entity is surprising since time expressions are not difficult to identify. However, it again might be due to the high number of entities, which is especially noticeable for time expressions.

Relations The relation annotation was not done on gold standard entities. Therefore, the underlying entities marked by either annotator might not even be the same in some cases, as is evident by the entity agreement scores. However, the most important relation for our use case, namely a `caused` relation between `drug` and `disorder` mentions still gets the highest F_1 score among the relation types with a score of 0.60. Note that for this relation, the annotators have an agreement of 0.63 and 0.72 with the gold standard.

Another interesting relation is the `has_dosage` relation between a medication and a measure. There is almost no agreement between the annotators, but when calculating the agreement with the gold standard, we see scores of 0.69 and 0.07, meaning, most probably, that annotator 2 simply forgot to annotate this relation. Further, the annotation of the relation `has_result` between a test and a measure is missing in Table 4.8, that is, the annotators did not use this relation, although there are relevant constructions in the text and therefore in the curated annotations.

Another example of incomplete annotations is the `has_route` relation which results in a IAA of 0.07. When comparing to the gold standard, the annotators achieve scores of 0.54 and 0.20, showing that they apparently annotated some of the relations existing in the gold standard, but not the same ones the other annotator did.

`is_opinion_about` seems to be another difficult relation where none of the annotators was successful. Their scores remain low even when compared to the gold data (0.15 and 0.20 for the relation between `opinion` and `drug`, respectively).

For the relation `signals_change_of` between a trigger and a medication mentions, the F_1 score is rather low, too. We attribute this to the differences in the trigger annotations, which are arguably difficult and often allow to choose from several potential triggers. This, in turn, leads to different entities being connected with the drug mentioned. The scores of the annotators for this relation when compared to the curated data are 0.45 and 0.29, respectively.

Finally, the relation `treatment_for` also shows a specific behaviour: The IAA for this relation between a medication mention and a disorder is 0.41, whereas the scores for the annotators when compared to the gold data result in 0.44 and 0.75, showing a big difference between annotators. On the other side, the same relation between a medication and function shows no agreement at all between annotators, but a closer look reveals that annotator 1 barely used this relation ($F_1 = 0.070$) while annotator 2 used it quite frequently and agrees with the gold data with a score of 0.53.

With respect to relations, we found that there were some relations completely ignored by one annotator, while some others were ignored by the other annotator. This leads to the question why this is the case. Several possibilities that could be improved come to mind, including the time spent on training the annotators, the high number of entities, relations, and attributes, and the annotation process, in which all levels were simultaneously annotated. Further, the thoroughness of the annotators might have played a role, aggravated by the aforementioned issues. This again demands a way to validate the thoroughness of annotation *while annotating*. For instance, a checklist at the end of each document might help the annotators to be reminded of every potential relation they could possibly apply.

Attributes Regarding attribute annotation, only the negation attribute shows a at least mediocre score of agreement. All others are rather low. When looking at the agreement between the annotators and the curated data, we find that the overall F_1 score for annotator 1 is 0.59 while it is 0.47 for annotator 2. Again, there are strong differences between the annotators. For example, annotator 1 agrees with the gold standard on the time attribute with an F_1 score of 0.71, but annotator 2 only agrees with a score of 0.53. This is reversed for negation, where annotator 1 reaches an agreement of 0.49 and annotator 2 has an agreement of 0.61.

The low agreement of the opinion attribute is difficult to justify since we found during curation that it was rather easy to decide on the sentiment of an opinion. However, since there was already a low agreement on the spans describing an opinion, this also reduces the chance of two annotations agreeing on the attribute for this opinion. For example, annotator 1 very rarely used opinion as an entity type and therefore agrees with the gold standard only with a score of 0.07 while annotator 2 used the opinion type more frequently and agrees with the curated data with a score of 0.37.

Lastly, we also see an almost non-existent agreement on the drug attribute, although the drug entity type itself shows a high agreement. Taking a closer look, we find that annotator 1 agrees with the gold data with a score of 0.60, but annotator 2 only agrees with 0.14. When taking a look at the annotations, we found that these attributes were simply not annotated, contrary to annotated falsely.

Summary In general, as mentioned above, we found during curation that neither annotator conducted "wrong" annotations (only very few), but that the combination of both annotators' annotations resulted in a more complete picture. We further observed high discrepancies between annotators and specific annotation types, where one annotator performed well on one type but not on another (i.e., the type was not used at all or very rarely) and vice versa for the

second annotator. This is unfortunate since it increases the curation effort. It also highlights the difficulty of annotation and the need for a method that helps the annotators review their own annotations.

Possible improvements could be supported by a more thorough pre-annotation (we only provided drug and disorder annotations), but also by a third and maybe fourth annotator simulated by a large language model like `Llama` (Touvron et al., 2023). A model like that could be fine-tuned with a few human-labeled examples and used as a complement to human annotators. Although we should not trust an LM alone, especially not in the medical domain, it could still provide a consistent annotation, balancing the inconsistency of human annotators. During curation, its predictions could then help to decide on final labels for, e.g., entities in a majority vote between human and LM annotators, reducing some of the curation load and also avoiding the cost of a third or fourth annotator.

Final Dataset annotated with Entities & Relations

In this section, the datasets as of the time of writing are described. Note that we aim to develop the data further, meaning adding more documents with annotations, but also different annotations, particularly for concept normalization. We show statistics for German and French next to each other to allow for a better comparison. An overview of the general statistics is provided in Table 4.10 and the number of annotations is given in Table 4.11.

		#tokens				#sentences			
	#docs	**total**	**mean**	**max**	**min**	**total**	**mean**	**max**	**min**
de	118	29,032	246.03	815	55	1,674	14.19	50	1
fr	100	18,184	181.84	463	42	969	9.69	25	1

Table 4.10: An overview of the currently annotated data in German (de) and French (fr). It shows the number of documents for each language, the total number of tokens and sentences, as well as the mean, minimum, and maximum number of tokens and sentences per document. The documents were split into sentences and tokens using `spacy`.

	entities			relations			attributes		
	total	**types**	**mean**	**total**	**types**	**mean**	**total**	**types**	**mean**
de	3,487	12	29.55	2,163	12	18.33	1,141	4	9.67
fr	1,939	12	19.39	1,129	12	11.29	537	4	5.37

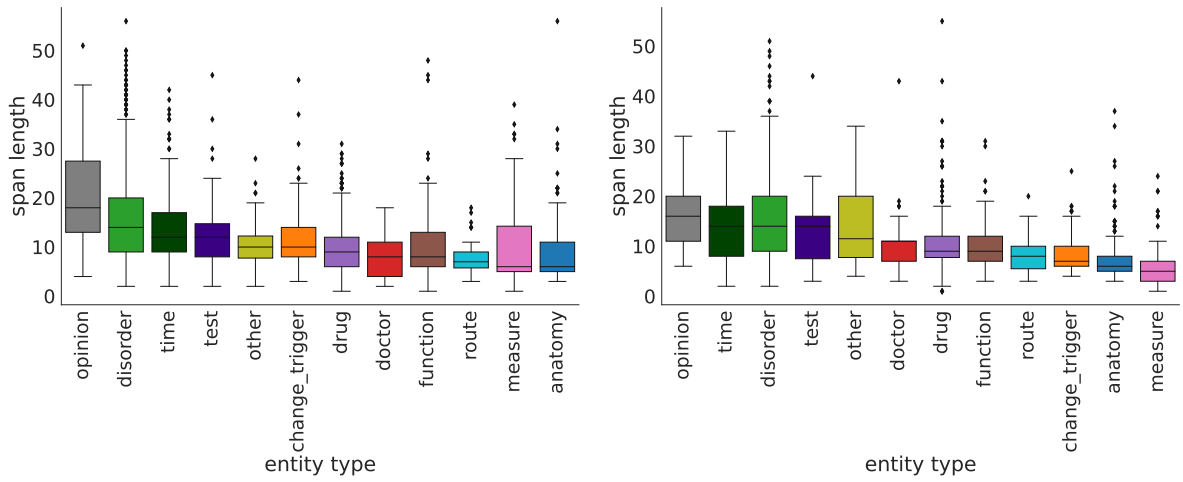
Table 4.11: Overview of the annotated entities, relations, and attributes for the German (de) and French (fr) data. It shows the total number of marked spans, the number of different types, and the average of annotated spans per document.

German We start with the German dataset, describing the first batch containing 118 documents. The German data contains about 29,000 tokens in total, with an average of 246 tokens per document. Some documents are rather long, with a maximum of 815 tokens, but most are between on average 100 and 300 tokens long (Figure 4.6).

Further, this part of the corpus contains annotations of 3,487 entities, distributed across 12 types as presented in Figure B.6. For German, the most frequent entity type is `disorder`, followed by `drug` and `time`. The maximum number of entity annotations found in a document is 90, while the smallest is 4. The span lengths of the single entities vary considerably, as shown in Figure 4.5: `opinion` and `disorder` are among the longest spans, with some spans

drug	disorder	translation
ads	Gelenkschmerzen	joint pain
arimidex	sehr schlimme Nebenwirkungen	very bad side effects
cerazette	3kilo runter	3 kilos down
estreva gel	vermehrte, starke Kopfschmerzen	increased severe headaches
mtx	Haarausfall	hair loss
opipramol	Watte im Kopf	"cotton in my head"
schmerztabletten	hauen mir die Schuhe weg	"knock my shoes off"
utrogest	wilde Träume	wild dreams
venaflaxin	Unwirklichkeitsgefühle	feelings of unreality

Table 4.12: A selection of ADRs as found when extracting the mentions annotated with `drug` and `disorder` and connected by the relation `caused`. Tokens were lower-cased.



(a) German data

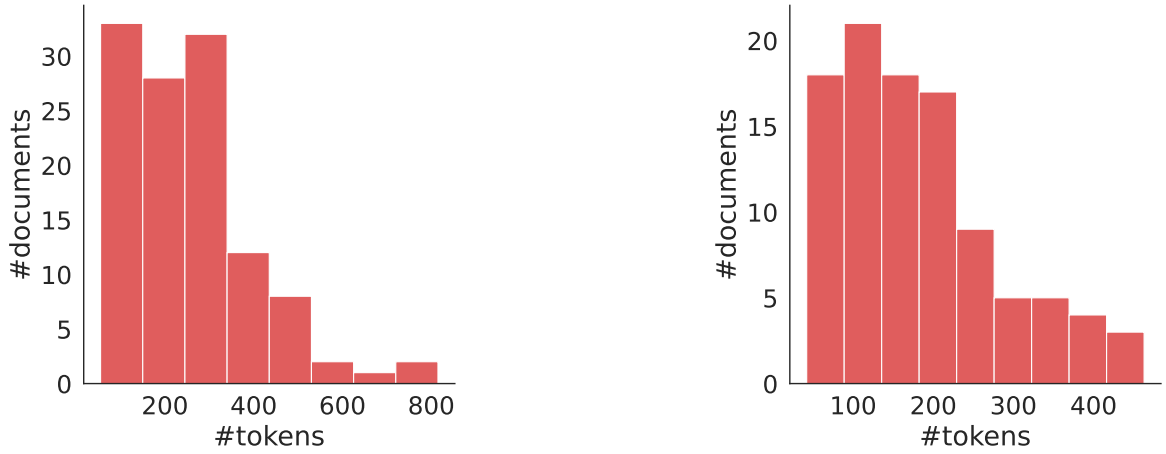
(b) French data

Figure 4.5: The distribution of entity span length (in characters) per entity type.

being longer than 50 characters. This is understandable, since the personal assessments of the patients can be very descriptive, and the same is true for descriptions of symptoms or signs. The latter shows a high number of extreme lengths. Note that `function` mentions tend to be shorter than `disorder` mentions. Mentions of medications can be found in the middle of the spectrum when sorting the span lengths by median value. Often mentioned drug names are, for instance "AD" (***de:*** *"Antidepressivum"*, ***en:*** *"antidepressant"*), or "MTX" (***de:*** *"Methotrexat"*, ***en:*** *"Methotrexate"*). The medication names that occur most often are "Utrogest" (29 times), "AD" (28), and "Progesteron" (22). The three most often mentioned disorders are "Angst" (29 times, ***en:*** *"anxiety, unrest, worry, apprehension"*), "Nebenwirkungen" (25, ***en:*** *"side effects"*) and "Schmerzen" (19, ***en:*** *"pain"*).

The relation with the highest frequency is, not surprisingly, `caused`, with 598 annotations (Figure B.7). It is followed by the `has_time` relation, which is also no surprise given the 468 mentions annotated with entity type `time`. Most of them (344) are indeed connected with a relation. The relation with the lowest frequency is `interacted_with`, which was only used once. In Figure B.8 the distribution of head and tail entities per relation type is shown as well.

Finally, for the German data, the entity type with the highest number of attributes is `time`, which was assigned more than 600 times. Of this, most values fall into the categories `duration` and `point in time`. The second highest number of values has the entity type `opinion`,



(a) German data

(b) French data

Figure 4.6: The distribution of document length of the German and English data. Note the different scaling on the axes.

which is mostly used for positive assessments of either drugs or (passed) disorders.

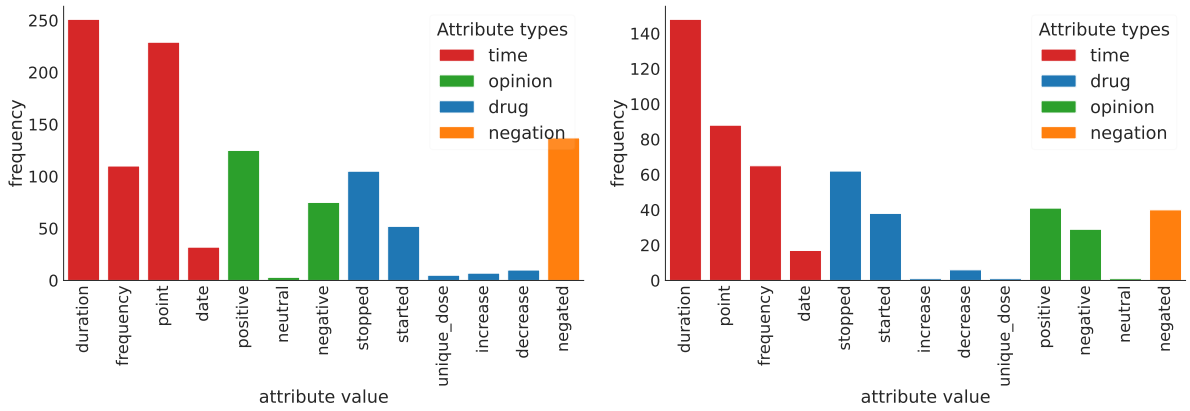
Table 4.12 shows some randomly selected ADRs as determined by finding `drug` mentions that are connected with a `disorder` mention via the `caused` relation. The complete table containing all 369 ADRs is shown in Table B.7.

French The French corpus currently contains 100 documents with about 18,000 tokens in total (Table 4.10). On average, each document contains about 181 tokens, much less than the German data. The majority of the French documents are about 50 to 200 tokens long.

Also, there are fewer annotated entities in the French data compared to the German part of the corpus, with 1,939 across documents using the same 12 types. As for the German data, the most frequent type is `disorder`, again followed by `drug` and `time`. Then, the order changes when compared to the German data, with `test` being the entity type with the lowest frequency. The most often mentioned drugs are "progestérone" (31 times), "utrogest" (13), and "gynokadin" (11). Note that these most likely have different names in French (except for progesterone, which is a hormone), but we kept the German names for simplicity. The disorder with the highest frequency are "nausées" (16 times, ***en:*** *"nausea"*), "vertiges" (11, ***en:*** *"vertigo"*), and "diarrhée" (8, ***en:*** *"diarrhea"*).

Like German, `opinion` shows the highest median in length compared to the other spans. This is followed by `time` and `disorder` entities. Interestingly, mentions describing trigger words for drug changes seem much shorter in French than in German. Note that in the French translations, we did not use abbreviations, but translated the written-out expression, since we do not know what kind of abbreviations are used in French patient fora. This might explain, for example, the differences in the span length distributions of `drug` and `doctor` mentions. Of course, these are also due to general differences in the two languages.

Regarding relations, the French data is again similar to the German data, both show the highest number of annotations for `caused` (598 versus 342 times) and `has_time` (344 versus 270). Again, `interacted_with` only occurs once. However, the use of the `is_opinion_about` relation seems to be much less frequent relative to the other relations when compared to the German data.



(a) German data

(b) French data

Figure 4.7: The distribution of attribute values for each attribute type: `time` (duration, frequency, point in time, date), `opinion` (positive, neutral, negative), `drug` (stopped, started, unique dose, increase, decrease) and negation (yes/no).

Similar to the German part of the corpus, the `time` entity receives the most attributes, namely 318. Most of these are again `duration` attributes, with `date` showing the lowest number of assignments. The entity type with the second most attributes is `drug`, where most are set to `stopped`. Again, the `opinion` attribute is dominated by positive assessments. Altogether, we can extract 313 ADRs, all of which are shown in Table B.8.

4.1.4 Limitations of the Corpus

As with any other corpus, the presented one has limitations as well. First of all, although we randomly sampled the posts from the Lifeline forum, a huge majority are written by women in a specific age range, judging from the topics of the posts. In general, it seems more likely that patient fora attract people of a certain age since younger users tend to be active in other networks. Therefore, only a small part of the population is represented, similar to corpora created from Twitter or Reddit posts. Nevertheless, it can still complement other data gathered for extending the knowledge about ADRs.

The same is true for the data source of the German and French corpus: Currently, all data originates from only one forum. Having data from multiple sources would make the corpus more representative and most probably would cover other occurrences of ADRs. This is part of future work, since we now have the means to train models on the newly provided data, making it much easier to extend the corpus both in terms of more data and annotations. Further, having more French data, particularly original texts that are not translated, would improve the quality and generalizability of the corpus.

Social media data also come with the caveat that the claims, experiences and advise users post might not necessarily be true – out of a lack of knowledge, ignorance, or purposefully, to misguide other patients. This needs to be taken into account when using our corpus, although we removed speculative content and aim to annotate these samples in future work. In general, the extracted information and annotations should not be used in any medical application before consulting a pharmacovigilance expert.

Similarly, although we de-identified the data, there might still be remaining traces of users. It therefore needs to be noted that it might still be possible to identify the users since the fora are publicly accessible. However, our corpus will be distributed only via a data protection agreement and only within the research community.

Another limitation, or rather a challenge, is the high imbalance in document labels for the data annotated with binary labels. This is normal for this kind of data but makes automatic processing more intricate as we will see in Chapter 5.

The other levels of annotation, that is, entities, relations, and attributes, can be improved as well. The annotators were asked to use the wildcard entity type `other` and relation `misc` in case they found relevant information for which a type did not exist yet. Both were frequently used several times during annotation, demonstrating that our pre-defined set of entities does not cover some expressions relevant to the health information. This includes, among other things, entity types describing therapies or other medical terms. With respect to relation types, we found that some more would have been helpful, e.g., relations connecting drug mentions and anatomy mentions and relations leading from an opinion to other entity types like `route` and `anatomy`. We also found that some disorders were caused by the route of a medicine, not by the substance itself, a case we did neither consider nor encounter when testing the annotation scheme.

Further, expressions often annotated in EHRs are so-called social determinants of health (SDH) (Lybarger et al., 2023), that is, "non-medical factors that influence health outcomes"[29]. SDH might be more frequent in EHRs, but even in the forum texts, we found many factors patients mentioned which either increased ("Ich mache gerade KG und es bekommt mir sehr gut.", ***en:*** *"I am currently taking physiotherapy and it works very well for me."*) or decreased ("Bruder hat Gehirntumor, Tod unseres Hundes", ***en:*** *"Brother has brain tumor, passing of our dog"*) their health status. Annotating these might give a more complete picture of a patient's health and the factors people struggle with when taking medication.

Finally, we also cannot rule out the existence of annotation errors, even after curation. As shown by the IAA scores, all types of annotation seemed to be rather difficult, resulting in annotation mistakes. Those could be reduced by better training of annotators and the above-mentioned "checklists" to help annotators be more thorough. Also, additional annotators based on LMs could help to improve annotation.

4.1.5 Summary and Conclusion

In summary, with the KEEPHA dataset, we provide a corpus of UGTs with a focus on adverse drug reactions. It is unique in its combination of domain (social media), language (French, German, Japanese), and annotation (binary, entities, relations, attributes). We showed that it is possible to create a multi-lingual dataset where each language's part is based on the same guidelines (RQ 1). The part of the corpus presented in this thesis is in German and French and originates from a patient forum. We double-annotated and consolidated 10,000 documents with binary annotations including 324 describing ADRs in German. A smaller amount of these texts were translated into French, comprising (currently) 100 positive and 764 negative documents.

Out of those, the positive documents were further annotated with entities, relations, and entity attributes, resulting in 118 German and 100 French documents, with 3,487 and 1,939 entity annotations, respectively. Annotation is ongoing and we are currently also annotating negative documents that do not contain ADRs but all other entity types. With this, we provide a new corpus on a topic much needed for pharmacovigilance and public health research in languages other than English.

In contrast to related work, we developed annotation guidelines with the goal in mind to make them applicable across languages, even to those from different language families. We showed this by applying the guidelines to the languages German, French, Japanese, and English. Different from other non-English corpora on ADRs, our corpus does provide very

[29] https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1

fine-grained annotations. Also, we define an ADR by using relations, not entities. All signs, symptoms, and diseases are marked as `disorder` and only by relating them to a `drug` entity using the `caused` relation, we label them as ADR. In total, we identified 369 ADRs caused by different medications for German. In the future, this corpus is going to help in the extraction of more detailed information from UGTs.

To provide more context and allow relations across sentence boundaries, the documents are not split into sentences. This results in quite large documents, but also allows a more detailed perspective on health-related issues. Further, the data are not filtered by drug mentions, as most other related corpora are. In contrast, we take the label and entity distributions as are, resulting in a high number of negative documents but also providing the real data distribution with a wide variety of mentions, also with respect to disorder descriptions.

The scores calculated for IAA clearly reflect the difficulty of the task: Even binary annotation in this domain seems to be very difficult, due to the ambiguity of UGT, medical inconsistencies of patients describing their ailments, and very long documents with implicit mentions of ADRs. In this, we see parallels to other corpora based on user-generated content but also room for improvement for both annotation guidelines and annotator training.

Finally, the annotation guidelines are not only applicable to texts containing ADRs but also to other health-related topics written from a patient's perspective. Also, they should be applicable to other sub-domains in biomedical texts. Some entity types might be excluded, such as `opinion`, but most should be adaptable to any other type of text.

For future work, some directions were already mentioned. Although already extensive, the number of annotation types could be extended. Social determinants of health might be another path to follow as well as the (already planned) concept normalization, particularly for disorder mentions. So far, not many works exist on the automatic normalization of user-generated descriptions for languages other than English. The curation process could also be improved, for example, with an LLM to add a source of consistency as opposed to the inconsistency often induced by human annotators. A third LLM-annotator could help in deciding on annotations via a majority vote. On a more application-wise note, it would be interesting to compare the resulting ADRs with existing databases and/or medical experts, to see if all of the ADRs found are already known.

4.2 User Privacy in Health-related Data from Social Media

Since a lot of works not only in the biomedical domain but also in many other domains of NLP collect and analyze user posts from social media platforms, it is clear that data originating from these networks are important resources. As described in Section 3.5, using these data comes with several issues that need to be considered, especially to preserve the privacy of the users but also to avoid infringing copyrights. In the following, two methods with the goal to circumvent these issues are proposed, i.e., how to use social media data without hurting people's privacy and right on data protection. The first method deals with the question if it is feasible to directly ask patients for their data (Section 4.2.1), while the second one is a teaser to the NTCIR'17 MedNLP shared task on social media (Section 4.2.2) and the lessons learned when generating synthetic tweets.

4.2.1 Collecting Sensitive Data with Users' Consent

Biomedical and clinical NLP is based on very specific data, e.g., texts that contain patient-relevant information like their medical and social background, their age, or gender. Some datasets also need to be associated with a certain disease or drug. In Section 3.2 we already reviewed some examples. However, obtaining these data takes place in a rather grey area:

For most platforms, it is technically allowed to download data from the mentioned websites and users allow for their data to be downloaded and re-used when they agree to the terms of service. But since terms of agreement or data protection regulations are not the easiest to read and understand by non-professionals, users often just click "agree" to be done with it. Therefore, although users choose to post messages about their health on Twitter and other platforms, they probably never actively and consciously consent(ed) to these messages being processed and, maybe more importantly, being published as datasets by researchers.

This poses a problem in NLP: If a dataset cannot be distributed such that other researchers can use it for their own work, it does not have much value. Therefore, either the data is distributed anyway (that is, without the users' explicit consent), or researchers refrain from using the data, although they would be such a valuable and insightful resource. An obvious option to circumvent this dilemma is to actually *ask* the patients for their consent to distribute data. Note that this mainly concerns social media data like Twitter, Facebook, and other openly accessible pools of information. For other resources, like EHRs, researchers should have consent to process the data from the beginning. However, asking users on the mentioned platforms is, again, very time-intensive and difficult to implement.

In the following, we provide a prototype for a qualitative study to check the feasibility of gathering descriptions of patients in two languages on the topic of adverse drug reactions. For this, we designed a survey in which participants are explicitly asked for consent with respect to sharing their experiences with medications and ADRs. First, the research question and study design are described. Subsequently, the results of the study are presented and analyzed.

Methodology

The aim of the presented study was to collect written texts describing experiences of medical side effects from a patient's perspective, making it possible for users to actually *consent* to share their data. The research question we are addressing with this is RQ 2.

For the presented prototype, high quality in this scenario implies textual messages similar in style to tweets or forum posts and describe medical issues, as opposed to simple answers to drop-down menus often used in user surveys. To approach this question, we designed a survey that first enquires about some general circumstances with respect to the user and then goes deeper into detail by asking specifically about experiences with medication intake and possibly resulting side effects. In total, there were 15 questions to answer. Most of them were not mandatory or, if they were, they allowed a "prefer not to say" option, since we wanted the participants to feel as comfortable as possible while sharing their personal experiences. Also, we provided examples to show what kind of information we would like to see, mainly to demonstrate that informal answers using lay terms are allowed (and even preferred), too. Of course, this might also introduce a bias for the participants to respond similarly, but we decided to take this risk to get data similar to what we found in online patient fora. The main parts of the survey are as follows[30]:

Page 1 The participant is given some information about the project, the data privacy procedure, and a contact person's e-mail address. They are then asked to consent or not consent to participate in the questionnaire.

Page 2 After giving their consent, the participant is forwarded to the next page and asked to create a personal code that can be used for deleting their data if they ask for it after completing the survey.

[30]The concrete questionnaire: https://cloud.dfki.de/owncloud/index.php/s/ksaH7sSrDyR867R.

Page 3 This page inquires about some general questions. One important aspect is if the participant is reporting about their own experiences or about another person's experience, e.g., their partner's. The other questions concern the participant's age and gender.

Page 4 With this page, the main part of the survey begins. Participants are asked to enter the medication they take in their own words via a text box, e.g., by just listing the drug names ("Ibuprofen, Aspirin" or simply "pain medication"). Following that, they are asked if the medication was prescribed or not. They can also leave additional comments. At the end of this page, the participants are asked to enter their diagnosis and medication dosage in two text boxes.

Page 5 This page is designed to get the information most important for the project. First, participants are asked to describe if they felt better or worse after taking the medication they mentioned on the page before. Then, they are asked to rate (approximately) the number of side effects they experienced on a Likert scale (Likert, 1932), ranging from "no side effects at all" to "a great many side effects". After that, they are required to enter a description of their side effects into a text box. Finally, they are asked if they will continue using the medication and if they would recommend it to other patients. A final text box allows them to enter additional comments relevant to their experience.

Page 6 The last page thanks the participants and repeats the contact information in case of questions. If desired, participants can enter feedback for the survey.

The questionnaire was prepared in both English and German to allow for a bigger pool of participants and a comparison of the given responses in future work. We used the survey provider LimeSurvey[31] hosted on servers in France[32]. It was distributed via the survey platforms SurveyCircle and SurveySwap[33], but also via the social media platform Reddit.

Both survey distribution platforms are based on earning credits by responding to other users' surveys. The more credits, the higher the survey is ranked among all surveys for a specific region (SurveyCircle) or the more participants are likely to see the survey (SurveySwap). Both platforms are free to use but require registration.

For Reddit, we mainly posted the survey URL(s) in so-called subs (as in sub-threads) "SampleSize" (English and German), "samplesize_DACH" (German-speaking countries, i.e., Germany, Austria, Switzerland), "MenoPause" (mostly English), and "MedicalQuestions" (mostly English) and as a comment to the poll thread in the sub "Wissenschaft" (mostly German). We chose these subs since they allowed the posting of surveys in general (in some subs it is forbidden completely) and had some medical associations, mostly people talking about specific medical issues. Since new surveys are posted frequently, we had to re-post the survey several times, approximately every three days starting from February 13, 2023.

After running the survey for one month, we downloaded the results from LimeSurvey. We summarized the participant statistics (number of participants, used platforms, languages, age, etc.) and qualitatively evaluated the responses to the text boxes. The findings are described in the next section.

Results

In total, the survey was accessed by 66 participants over the course of four weeks. However, not all of them completed the survey. We define a survey as completed when at least page 5 was

[31] https://www.limesurvey.org/

[32] The servers are maintained at the Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), CNRS, Université Paris-Saclay, in Orsay, France.

[33] https://www.surveycircle.com/en/surveys and https://surveyswap.io

answered since page 6 only contains the "thank you" message and the respective SurveyCircle or SurveySwap codes. We further excluded "completed" surveys where participants did not consent to share their data, which was the case for 12 participants. After filtering out non-consenting participants, we arrived at 54 responses.

In total, 33 participants responded to the German (de) version of the survey, while 21 responded in English (en). Figure 4.8 shows that out of the 54 participants, 3 left the survey after page 5 and 27 reached page 6. Therefore, we arrive at 30 final responses of which 16 were completed in German, while 14 were completed in English. Note the relatively high number of (German) participants leaving the survey after finishing page 3 in Figure 4.8.

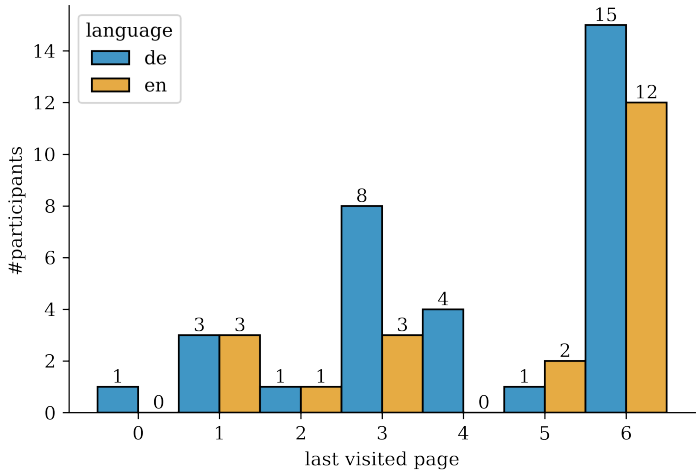


Figure 4.8: The number of participants per last page accessed, separated by language.

With respect to the dissemination platforms we used, we find that SurveyCircle and SurveySwap attracted the most participants (see Figure B.11). For Reddit, we can only see one occurrence, however, Reddit users could choose between the direct link and the detour via SurveySwap and SurveyCircle. We further observe a relatively high number of drop-out participants on page 3 for the direct link.

When looking at the age of the participants (they were asked to provide their birth year, but they were not obligated), we can see that most participants were born between 1975 and 2000. The age distribution is shown in Figure B.12. However, since some provided a birth year after 2020, they might not have been totally truthful.

Another interesting finding is the distribution of the number of adverse drug reactions. Participants were asked whether they experienced adverse reactions when taking medication and if yes, how many there were. Eleven participants answered with "a few side effects" (see Figure 4.9), while only two experienced "a great many side effects".

Finally, we plot the number of tokens in the given texts from the text boxes in Figure 4.10. The actual texts written by the participants are the most interesting for further NLP research. Therefore, we combine the boxes' content for medications, experiences, diagnoses, dosages, a more detailed description of the experienced side effects, and recommendations to other patients and tokenize the texts simply by white space. We find that the German texts tend to be longer and count, in total, 1,185 tokens for German and 553 tokens for English.

We further split the content of the text boxes into sentences[34] to get a closer look at the actual descriptions. By semi-automatic inspection, we find 45 unique drug mentions in the text boxes for medications. Some of them are very specific ("retardiertes Amphetamin", ***en:*** *"slow-release*

[34]We split strings by full stops and new lines.

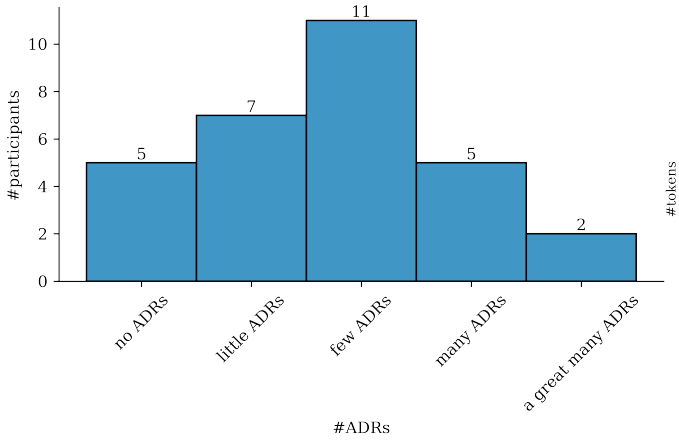


Figure 4.9: The number of ADRs as described by the participants.

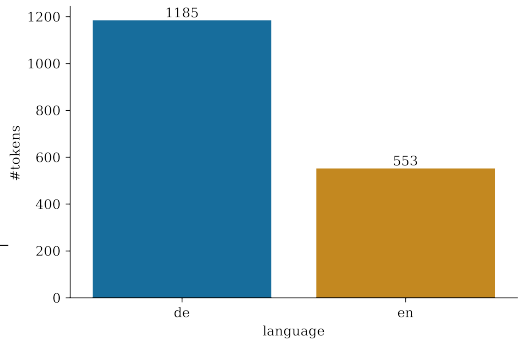


Figure 4.10: The number of tokens per language, summarized from the questions where free text was required (text boxes).

amphetamine"), while some are more generic ("doctor's prescriptions for flu (combination of painkiller)"). This is similar to the medication dosages: In the collected data, we count 33 descriptions of dosage (for both languages). Participants use, for example, expressions like "2 to 3 before bed", "2-3 Tabletten täglich" (***en:*** *"2-3 pills daily"*), "20mg a day", "2500 mg / tag" (***en:*** *"2500 mg / day"*).

The participants were also willing to share some of their diagnoses. We gathered 31 descriptions, ranging from headache to depression. Some preventive medicine was mentioned as well. Here, a change in writing style compared to the other text boxes is evident: Describing the diagnoses is mostly done by listing the issues, separated by commas, but not formulated in complete sentences. This applies to both languages.

When the participants were asked about their general experiences with medication intake, they described in more complete sentences than when asked about their diagnoses. The same is true for the description of ADRs. The patients seemed to try to describe the side effects as clearly as possible, exactly what we wanted to achieve. We count approximately 50 description sentences for general experiences and 40 sentences for adverse reactions.

Discussion

The described survey was designed to be as simple as possible. However, since we wanted to investigate whether it is feasible to collect natural textual descriptions using a survey, quite a lot of text boxes were added to the questionnaire. This induces a higher "workload" for participants, making it less likely for them to respond. Indeed, one participant even responded by saying they did "not want to type this much". Based on the relatively high number of participants that responded until pages 3 and 4, we also assume that they might have been demotivated as soon as they saw the text boxes. Nevertheless, the texts were quite long and detailed, much more so than expected. Also, these texts were (mostly) rather meticulous descriptions of health issues and their consequences, and not simple lists of drug names etc. To further improve the outcome, the questionnaire could be made more interactive, as in a chat-bot scenario, or maybe also accessible via voice messages, which many people prefer over typing.

Another critical point in a survey about medication intake is the sharing of very personal information which might make some patients shy away from participating. However, although some participants did not consent and left the survey, we still received a good number of responses, exceeding our expectations.

A survey takes much longer and returns less data than other data collection methods. Also, the repeated posting on various platforms took some time. This, however, might be handled automatically for future work, and data collected like that also comes with some simple annotations attached, based on the text boxes the participants respond to. Note that the survey was disseminated only via a few platforms which are a significant part used by younger people. The survey platforms SurveyCircle and SurveySwap are mostly visited by students who also need participants to respond to their surveys. This restricts the pool of participants and needs to be kept in mind when interpreting the results.

Although the amount of data collected via the survey (25.9 kB) is not even close to the number of samples automatically scraped from the web, we found that the texts were of good quality for our purpose, which is using them in biomedical text processing of user-generated text. With this, we do not refer to "grammatically correct" texts, but rather to texts that were written *by laypersons to laypersons*, using a mixture of languages and a range of expressions for similar semantic outcomes (see, for instance, the descriptions of dosages). By manual inspection, the responses seemed to be very similar to texts collected from Twitter and patient forums, displaying the variety in which people describe the things they care about or worry about. A more thorough comparison of the commonalities and differences between the different types of user-generated texts is, therefore, an interesting direction to follow.

Finally, the most important aspect of the presented pilot project was to investigate whether people would indeed share their personal experiences when asked for it. Contrary to what we expected, people actively made the decision to consent and provide answers. And even though we did not gather an enormous amount of responses, it still shows that it can be done. An approach like this might include more effort on the researcher's side, but it also supports *the right to be heard* from the patient's side. For (large) language models, some good quality and expressive examples might already be enough to fine-tune a model into the right direction (Brown et al., 2020), in this case, the detection of ADRs.

Conclusion

We conducted a prototype study for the feasibility of eliciting data relevant to NLP, and in particular, automatic ADR detection, by creating a survey and asking participants to describe their medical side effects in either English or German. One important part of this survey was that users were actively asked to consent to sharing the data, in contrast to how patient-related datasets in NLP are usually built.

Contrary to our expectations, quite a few people (30 usable responses in total) responded during the one month the survey was running. Although the amount of gathered data cannot keep up with the huge datasets built during the last years, the quality is good enough and might help in the few-shot transfer of multi-lingual ADR detection. A more thorough inspection of the collected data and a baseline model is needed to investigate this further.

For future work, the survey would need to be improved, maybe with the help of specialists. For example, one participant mentioned that it would have been helpful to have lists of drug names to choose from since they could not remember all medications they were prescribed. We were thinking about doing this when designing the survey, however, it would have either restricted the number of available medication names or become another burden for participants, because there would have been too many medication names to choose from.

Another aspect that can be improved is the balance between not explaining too much of the study's goal to avoid biases and too much text and explaining enough such that participants feel safe sharing their experiences. An associated website with more details on the exact research might help in this regard. Further, the survey should be distributed via more relevant channels and translated into more languages to reach more people. Currently, it is only representative of a small number of people, mostly students and Reddit users.

In summary, the prototype worked better than expected and has still potential to be improved in various directions. The actual effect of the collected data still needs to be investigated.

4.2.2 MedNLP-SC-SM Corpus

This section discusses an alternative to collecting and particularly *distributing* data without explicit user consent which is based on the creation of synthetic data. For the NTCIR[35] challenge on Medical Natural Language Processing for Social Media and Clinical Texts (MedNLP-SC)[36], organized by the NII in Japan, the Japanese team of the KEEPHA project decided to run a challenge on the detection of ADRs in social media texts, following their previous involvements in the NTCIR conference (Aramaki et al., 2014; Wakamiya et al., 2017; Yada et al., 2022). On account of our trilateral project, the rest of the KEEPHA consortium, i.e., the French and German team, decided to contribute as well, turning the initial bi-lingual task into a (potentially) multilingual one, covering the languages Japanese, English, French, and German. Since the task is still running at the time of writing[37], the following discusses the overall process for creating yet another but different multi-lingual dataset and some preliminary insights focusing on the lessons learned.

Data Generation

Disclaimer: I was not involved in the dataset generation, but report the procedure for completeness.

The main idea underlying the MedNLP-SC-SM data[38] is the creation of a synthesized dataset that is distributable without hurting users' privacy. For that, Japanese tweets were generated, annotated, and translated into English, French, and German. The exact procedure was as follows: First, Japanese tweets were collected from Twitter using the official Twitter API and 68 disease names from a diseases dictionary (Ito et al., 2018) as query terms. Following that, the tweets were annotated with medical entities, using the NER model provided by Nishiyama et al. (2022). Tweets for which the model could find no symptom mentions were discarded. The filtered tweets then served as fine-tuning material for the Japanese version of the `T5` model[39] (Raffel et al., 2020).

Using a pre-defined set of medications as seed terms, `T5` was subsequently used to generate 11,000 pseudo-tweets per drug. Post-processing then took care of removing duplicates, pseudo-tweets without mentions of drugs or symptoms, and pseudo-tweets that were too close to the original. This resulted in 10,000 tweets overall.

In the next step, all remaining tweets were manually annotated by in-house[40] annotators, who focused on positive and negative mentions of ADRs, drug names and general health-related complaints. The number of ADR mentions was counted and the most frequent 22 occurrences were used as labels for each pseudo-tweet. Thus, in the end, each pseudo-tweet was tagged with 22 labels which were either set to 0 (ADR symptom does not occur) or 1 (ADR symptom is present). The labels were furthermore mapped to UMLS CUIs.

[35] http://research.nii.ac.jp/ntcir/ntcir-17/index.html

[36] https://sociocom.naist.jp/mednlp-sc/

[37] The conference will be in December 2023.

[38] SM is short for social media and distinguishes the task from the MedNLP-SC-RR task about radiology reports.

[39] https://huggingface.co/sonoisa/t5-base-japanese

[40] In-house meaning the medical staff in the Social Computing Lab at NAIST; https://sociocom.naist.jp/index/

As a last step in the generation process, all pseudo-tweets were translated into the aforementioned languages using the translation service DeepL again[41]. Note that generated emoticons, emoji, and kaomoji[42] were removed as well since they were mostly nonsensical and did not translate well.

A Japanese pseudo-tweet and its translations are shown in Example 4.17. All four texts are labeled with the symptom *rash* (both "measles" and "rash" are mapped to the same CUI C0015230). All other symptoms are set to 0. The participants of the shared task are asked to submit a system that is able to predict if one or more of the 22 symptoms occur in a given example.

(4.17) a. ***ja:*** "アザチオプリンを服用して2ヶ月経ちました。1週間くらいで全身の発疹はなくなり、かゆみもほぼ無くなっていたのですが、麻疹が少し残ってて怖かったなぁと思います。"

b. ***en:*** *"I've been on Azathioprine for 2 months now, and after about a week the rash all over my body was gone and the itching was almost gone, but I still had a bit of measles and I think it was scary."*

c. ***fr:*** *"Je prends de l'azathioprine depuis deux mois maintenant, et après environ une semaine, l'éruption cutanée sur tout mon corps avait disparu et les démangeaisons avaient presque disparu, mais j'avais encore un peu de rougeole et je pense que c'était effrayant. "*

d. ***de:*** *"Ich nehme jetzt seit zwei Monaten Azathioprin, und nach etwa einer Woche war der Ausschlag am ganzen Körper verschwunden und der Juckreiz fast weg, aber ich hatte immer noch ein bisschen Masern, und ich glaube, das war beängstigend."*

Data Validation

Disclaimer: I was not involved in applying the different validation measures, only in interpreting their results, and report the procedure for completeness.

During the generation process, two factors were introduced that had the potential to reduce the comprehensibility of the pseudo-tweets: (i) generation, and (ii) translation. The latter only applied to the English, French, and German data. Through the translations, wrong information could have been inserted, rendering the labels invalid. Further, the pseudo-tweets might not be understandable anymore. Thus, to improve the dataset's quality, the team validated the data to find "suspicious" pseudo-tweets, i.e., those that were noticeably different from the original Japanese pseudo-tweets or the other translations. For this endeavor, the following metrics were considered[43]:

Length ratio (Gale and Church, 1993): This measure counts the number of characters in a Japanese pseudo-tweet and sets it in proportion with the number of characters in the translations. For a better comparison, the Japanese pseudo-tweets were transliterated into Latin script using an open-source library[44]. The mean and standard deviation from the underlying normal distribution were calculated for each length ratio. Those pseudo-tweet pairs whose length ratio was outside the bounds of the respective normal distribution (for each language pair) were flagged as outliers, resulting in 172 (English), 284 (German), and 279 (French) outliers.

[41] The Social Computing Lab had a subscription for DeepL.

[42] Constructions from Japanese (and other) characters, e.g., ヽ(•‿•)ノ .

[43] For more details, we refer the reader to the NTCIR'17 MedNLP-SC-SM overview paper which will be released in December 2023.

[44] https://pypi.org/project/pykakasi/

Semantic Similarity Using LASER embeddings (Schwenk and Douze, 2017) of the given pseudo-tweets, the cosine similarity between the Japanese source and the translations was calculated. Pseudo-tweets falling below a pre-defined similarity threshold (per language) were flagged, resulting in 292 (English), 313 (German), and 306 (French) outliers.

Token Alignment The alignment between tokens of the Japanese source pseudo-tweet and a translation was calculated using SimAlign (Jalili Sabet et al., 2020). Then, the translations are flagged if the proportion of aligned tokens is below a pre-defined threshold per language. This procedure results in 110 (English), 88 (German), and 311 (French) outliers.

Back-translation + Token Alignment This repeats the procedure described before but uses the back-translated pseudo-tweets in each translation pair, that is, the Japanese source pseudo-tweets are aligned with the back-translated pseudo-tweets, which are now also in Japanese.

The resulting number of outliers are again shown in Table B.9 to compare the outliers per measure. Having found potential outliers, the next steps involved determining those texts where at least three out of four measures resulted in flags (see bottom row in Table B.9). The number of flagged samples overlapping across languages was determined to be 19. Then, some team members with the respective language skills[45] checked the flagged outliers.

After the first validation round, the data was used to train baseline models for the multi-label classification task. Applying the trained models to the data highlighted some inconsistent and difficult pseudo-tweets that were corrected or removed if needed. Finally, each language subset contained 9,957 tweets, i.e., 38 tweets were removed. All other tweets were reformulated to more accurately fit the original Japanese pseudo-tweet and the annotated labels.

Analysis

Other inconsistencies in data created like that are more difficult to find. Below, we highlight some examples with respect to translation (in-)consistency across languages and comprehensibility as well as the authenticity of the pseudo-tweets. These pseudo-tweets were encountered during validating the German tweets.[46]

Inconsistent translation across languages In Example 4.18, first of all, two symptoms are described, but only *nausea* is a side effect the person is actually experiencing. *Dizziness* is just stated as a general side effect of, probably, minocycline. Regarding the translation quality and consistency, compare, for example, the use of tenses in the translations. The English and French versions are more similar in their meaning than the German version. In English and French, the meaning of the sentence is "nausea is *currently* getting better, although I am *only* taking gastrointestinal medicine", while the German version rather states that the person "*had* nausea and it *only got* better using gastrointestinal medicine". The Japanese pseudo-tweet is more consistent with the English and French reading. This is due to the fact that in Japanese, there is only one past tense, while there are several ways to describe past events in the European languages. A similar phenomenon is happening in Example 4.17. Further, the mention of "measles" is confusing in Example 4.17, but this might be due to the generative model producing something that is similar to the first mention of "skin rash" (***ja:*** "発疹"), referring to it with "measles" (***ja:*** "麻疹").

[45] I was responsible for the English and German tweets and checked some of the French tweets.

[46] Many thanks to Prof. Ryo Nagata and Tomohiro Nishiyama for helping me verify the Japanese pseudo-tweets with respect to their translations!

(4.18) a. ***ja:*** *"<user_name> そうなんですね副作用でめまいとかあるんですね...。私の場合は、ミノサイクリンの副作用に嘔気がありましたが、カロナールと胃腸薬だけで良くなってきました～!ありがとうございます!!"*

b. ***en:*** *"<user_name> That's right, dizziness is a side effect... In my case, I had nausea due to the side effects of minocycline, but it's getting better with only caronal and gastrointestinal medicine ! Thank you!!"*

c. ***fr:*** *"<user_name> C'est vrai, les étourdissements sont un effet secondaire... Dans mon cas, j'ai eu des nausées à cause des effets secondaires de la minocycline, mais ça s'améliore avec juste des médicaments caronaux et gastro-intestinaux! Merci !!"*

d. ***de:*** *"<user_name> Richtig, Schwindel ist eine Nebenwirkung ... In meinem Fall hatte ich aufgrund der Nebenwirkungen von Minocyclin Übelkeit, aber nur mit Karonal- und Magen-Darm-Medikamenten wurde es besser ! Vielen Dank!!"*

Medical incorrectness Looking again at Example 4.18, we find that a "caronal" medication does not exist in either of the European languages. However, it seems to be a brand name of acetaminophen in Japanese. Therefore, while the Japanese tweet might be medically correct, this is not necessarily true for the translations.

Nonsensical / Self-contradictory tweets Example 4.19 shows self-contradicting texts across all languages. The person first states that the medication did not seem to work but then says its effect is very high. Note that this pseudo-tweet could also be read as "the medication did not seem to work, but in the end it did work", but at least in the German version that would have been written in a different way, i.e., it is not very authentic in the way it is written. We encountered several such examples, demonstrating that the generation process does not necessarily seem coherent and might contain incorrect medical expression.

(4.19) a. ***ja:*** *"<user_name> そうなんですね!!私の場合は、ミノサイクリンが効かなかったみたいで副作用にめまいとかあったりしましたお薬の効果はほんと高いですよ～!"*

b. ***en:*** *"<user_name> I see! In my case, the minocycline didn't seem to work and I had some side effects like dizziness. The effect of medicine is really high!"*

c. ***fr:*** *"<user_name> Je vois ! Dans mon cas, la minocycline ne semblait pas fonctionner et j'avais des effets secondaires comme des vertiges, mais le médicament est vraiment efficace !"*

d. ***de:*** *"<user_name> Ich verstehe! In meinem Fall schien das Minocyclin nicht zu wirken und ich hatte Nebenwirkungen wie Schwindelgefühl, aber das Medikament ist wirklich wirksam!"*

Incomprehensible pseudo-tweets There are also samples that simply do not make sense. In Example 4.20, English version, the pseudo-tweet gives the impression that minocycline is a side-effect and not a medication (***en:*** *"... I've seen it listed as a side effect of dizziness ..."*), it seems like the drug name and symptom were reversed. In both the French and German versions, this part is fine. However, the remainder of the translations are difficult to understand: Minocycline seems to be an antibiotic[47], and the patient first says that it has the side effect of dizziness, but then they say their headache is reduced and that they regret taking the medication. It is not clear if the minocycline caused the headache and why they regret taking it. The two Japanese pseudo-tweets in Example 4.19 and 4.20 are also incomprehensible.

[47] https://www.drugs.com/minocycline.html

(4.20) a. ***ja:*** *"<user_name> 私ミノサイクリン飲んでます♂副作用にめまいって書いてるのを見たことあるけど、私は飲まなかったなぁ～と思えるくらいには頭痛も治まりましたよー"*
b. ***en:*** *"<user_name> I take minocycline♂ and I've seen it listed as a side effect of dizziness, but my headache has gone away enough that I wish I hadn't taken it!"*
c. ***fr:*** *"<user_name> Je prends de la minocycline♂ et j'ai vu qu'elle avait pour effet secondaire des vertiges, mais mon mal de tête a disparu au point que je regrette de ne pas l'avoir prise !"*
d. ***de:*** *"<user_name> Ich nehme Minocyclin♂ und ich habe gesehen, dass es als Nebenwirkung Schwindel aufgeführt hat, aber meine Kopfschmerzen sind so weit weggegangen, dass ich mir wünsche, ich hätte es nicht genommen!"*

Vocabulary & Naturalness The pseudo-tweets in Example 4.21 can be read as self-contradictory again: First, the writer says they have no pain, then they describe it. However, it seems like the "no pain" (***fr:*** *"pas de doleur"*, ***de:*** *"keine Schmerzen"*) mentions refers to the "tubal angiogram" (this does not seem to be a correct medical expression, according to UMLS, this procedure is called hysterosalpingogram) the patient underwent, while the second mention of some kind of pain (***en:*** *"a sickening dull ache"*, ***fr:*** *"un mal sourd et désagréable"*, ***de:*** *"ein unangenehmer dumpfer Schmerz"*) refers to the experience after the treatment.

More interestingly, in the English and French versions, the translation uses two different words to describe the two experiences of pain, while the German one uses the same vocabulary for both and is therefore less specific. The English and German translations are unclear about what took "4 hours", either the "dull ache" or the "blood draws, IVs" etc. In the German version, the comprehensibility is even more confused with the relative clause "was etwa 4 Stunden dauerte" (***en:*** *"which took about 4 hours"*). The relative pronoun "was" refers to a neutral object; in the presented case, the only object with a neutral gender is "Eileiter-Angiogramm" (medical term: ***en:*** *"hysterosalpingography"*). However, according to the sentence structure, the closer object would be "Schmerz" (***en:*** *"pain"*), albeit a masculine noun. The French version, on the other side, refers with both descriptions of pain to the diagnostic procedure, while the sampling of blood, etc. took "about 4 hours". In the Japanese version, the second mention of pain is not directly perceived as pain, maybe similar to the English translation of "dull ache".

(4.21) a. ***ja:*** *"卵管造影検査、痛みなしただ気持ち悪めの鈍痛は続きましたね採血やら点滴やらなんやら4時間ほどかかりました <url>"*
b. ***en:*** *"I had a tubal angiogram, no pain, just a sickening dull ache that lasted about 4 hours of blood draws, IVs, and other things <url>."*
c. ***fr:*** *"Angiographie des trompes de Fallope, pas de douleur, juste un mal sourd et désagréable... La prise de sang et la perfusion ont duré environ 4 heures <url>."*
d. ***de:*** *"Eileiter-Angiogramm, keine Schmerzen, nur ein unangenehmer dumpfer Schmerz, was etwa 4 Stunden dauerte, einschließlich Blutabnahme und intravenöser Infusion <url>."*

Biases Finally, we also encountered what seemed to be language-specific biases. Japanese does not use pronouns the way French, German, and English do, but the other way around it is not possible to express certain things without using pronouns or gender markers in the European languages, with French distinguishing between two (feminine, masculine), and German and English distinguishing between three grammatical genders (feminine, masculine, neutral). Example 4.22 shows a bias of the translation system with respect to the language. The English version translates the "genderless" pseudo-tweet into the experience of a "little boy", while the French and German version speaks of a female child without saying "girl". In German and French, the gender is only detectable via the pronouns (***de:*** *"Sie, ihr, ihr"* and ***fr:*** *"elle"*) and indefinite articles (***fr:*** *"une enfant"*). Note that the German version first refers to a "child", which

has neutral grammatical gender in German, and then changes to "sie" (***en:*** *"her"*), although this could have also been solved by using "es" (***en:*** *"it"*), not referring to any gender whatsoever. However, this would have rendered the pseudo-tweets sound very unnatural. In Japanese, the last sentence does not have any subject, and thus, the translator needed to choose one based on its training data. Furthermore, the English version refers to "you" in the last sentences, while the French and German versions refer to the (female) "child" getting better.

(4.22) a. ***ja:*** *"<user_name>* そうですね。うちは風邪をこじらせて肺炎になった子がいて、抗生剤とステロイドの飲み薬だけもらっています早くよくなるように祈っていますね"
b. ***en:*** *"<user_name> Yes, that's true. I have a little boy who got pneumonia from a cold, and he's only getting antibiotics and steroids to take. I hope you get better soon."*
c. ***fr:*** *"<user_name> Oui, c'est vrai. Nous avons une enfant qui a attrapé une pneumonie à cause d'un rhume, et on ne lui a donné que des antibiotiques et des stéroïdes à prendre, alors croisons les doigts pour qu'elle aille bientôt mieux."*
d. ***de:*** *"<user_name> Ja, das ist richtig. Wir haben ein Kind, das durch eine Erkältung eine Lungenentzündung bekommen hat. Sie hat nur Antibiotika und Steroide bekommen, also drücken wir ihr die Daumen, dass es ihr bald besser geht."*

Another kind of bias can be seen in how the translations imitate the "writing style" of the Japanese pseudo-tweets, which, in turn, copies the original Japanese tweets the generative model was trained on. For example, tweets collected for the SMM4H shared tasks in English and French were much shorter and represented a different style of writing. Two samples are shown in Example 4.23.

(4.23) a. ***en:*** *"that nap was on point.... cymbalta did that shit cuz i dont take naps...ever"*
b. ***fr:*** *"depuis que j'ai arrêté deroxat pour effexor j'ai perdu 3 kilos sans rien foutre et en mangeant peut être mm +"*[48]

Use of Emoji As noted, the emoji and kaomoji in the Japanese pseudo-tweets were removed before translating because they did not make sense – they were generated randomly. Their existence nevertheless shows the frequent use of these means of communication in Japanese tweets. However, apart from their artificiality, even if they were valid emoji, they could and can not be directly translated into the European languages, since they often mean something different or are not used at all in those languages.[49] This is corroborated by Bai et al. (2019) (amongst others), who summarize research on emoji in different scientific fields and describe the influence of different cultures on the use and meaning of emojis.

Sentence Boundaries The pseudo-emoji in the pseudo-tweets could have also served as sentence boundary markers, since all examples except Example 4.17 lack these. Using emoji as sentence markers is very common across many languages (Sakai, 2013; Spina, 2019; Song, 2022).[50] We noticed, when using the browser version of DeepL, that the translation changed depending on how we presented the sentences to the system, i.e., split up as one sentence per line or the entire text, without sentence boundaries, as one. Sometimes, entire sentences were missing from the translations, another sign that DeepL might have problems with missing (sentence) boundaries. Conversely, having correct meaningful sentence boundaries in the Japanese pseudo-tweets might have increased translation performance and consistency.

[48] ***en:*** *"since i stopped taking deroxat for effexor i've lost 3 kilos without doing a damn thing and eating maybe even more."*

[49] Take, for example, the folded-hands emoji, which has, for instance, different meanings in Japanese and German (Western-Eurpean?) culture. https://emojipedia.org/folded-hands#emoji

[50] See also these slides: http://www.fluxus-editions.fr/grafematik2022-files/SONG-slides.pdf.

Summary

All the examples presented above are rather anecdotal. However, they still highlight the potential problems that come with the automatic generation and translation of texts not only but especially in the biomedical domain. First of all, translation systems might induce very small alterations that change the semantics of a sentence, making it less fitting for the labels previously decided on. This problem might be mitigated by only labeling after translating, but this would increase the annotation effort and make the texts potentially not parallel in their labels anymore, if, for example, a text in German does not get the same labels as the supposedly same text in Japanese. Second, the tweets might not be medically correct. Whether this is actually a problem is debatable since it might even be a good way of augmenting data with rare or seemingly non-existing associations between drugs and symptoms to make systems trained on these data more robust by providing more diversity.

Further, some texts can be unintelligible to humans, or, as shown, self-contradictory. Again, following the same reasoning as above, this might not be an issue. A system might still learn to recognize a medication and its potential (adverse) reaction even if it is not stated clearly in some of the texts used for training it. Having ambiguous and diverse examples could, again, allow a system to be more robust. Similarly, having texts with a varying diversity of vocabulary might not be a problem, depending on the task the data are used for. In addition, the translations as seen above make the different language sets rather *comparable* than *parallel*, since sometimes, the automatic translation does not necessarily contain the same content as the original – often, because the generated tweet itself did not make any sense or the translation misses sentences, but also because some expressions cannot be easily transferred across cultures and thereby across languages. What might be a problem, however, is the "correctness" of the pseudo-tweets in terms of spelling. There is not a single typographical error to be detected in the examples above, very unlike messages on Twitter and on social media in general. Therefore, when using the presented data as training material for a system, this might cause a reduction in performance on real texts in case they contain any errors.

Lastly, but very importantly, the biases introduced through the translation might indeed create an amplification of stereotypes, a topic much discussed in the context of LLMs like `ChatGPT` and similar models (Bender and Friedman, 2018; Talat et al., 2022; Névéol et al., 2022; Navigli et al., 2023, and others). If, for instance, the English examples always refer to a male person, while the German examples always refer to a female person, we should first investigate why this might be the case, and secondly post-process the data further to mitigate a potential reinforcement, e.g., by replacing pronouns.

In conclusion, the highlighted aspects demonstrate that generated and/or translated data should be handled critically, and not simply taken as is. Apart from "translationese", understandability and the transfer of labels might suffer, and unwanted artifacts, like biases or medical incorrectness, could be introduced to the data. Nevertheless, for data augmentation, these data might still be very useful.

Future Steps

Since the shared task is running, the above analysis only provided anecdotal evidence for the synthetic generation and validation of a multi-lingual comparable corpus. By manually inspecting samples of this new corpus, we found aspects that could be improved, and new research questions with respect to quality and usability of the presented corpus opened up.

Quality control *How can we further evaluate the quality of the presented corpus?*
We already described the first steps in validating and analyzing the corpus. However, a more thorough and systematic data validation is still necessary since even examples not marked as

outliers show specific characteristics setting them apart from real data. For example, the validation metrics described at the beginning could be extended with metrics similar to standard Machine Translation and summarization scores like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019). However, for those, we would need some kind of reference to compare with.

The metrics used mainly focused on the *translation* quality, not on the general quality of the texts. Regarding the generation part (which also concerns the translated tweets), an automatic evaluation is complex, since there is, again, no reference tweet with which the generated pseudo-tweet can be compared. One possibility would be to extend the model-based approach described earlier to a more difficult task like NER, as, for example, was done by Frei and Kramer (2022) and Hiebel et al. (2023a). They provided generated and then annotated data, trained a system on these data, and finally applied this system to a gold standard dataset which could be automatically evaluated. However, this would require more annotation effort. Another possibility would be the reverse, reducing the manual effort: Training a system on gold-standard data, e.g., the KEEPHA dataset, automatic annotation of the NTCIR data, and finally a manual evaluation of samples by humans.

For a more manual approach, we plan to sub-sample each language subset and manually rate the samples according to criteria such as naturalness, medical correctness, coherence, and fluency, independent of the other languages. The single-language approach makes it easier for the team since no one speaks all four languages well enough, and it helps to concentrate on the quality of the pseudo-tweets itself and its associated labels, without "distraction" from the other translations. First, however, the criteria need to be defined clearly to allow a simple yes/no evaluation[51]. Furthermore, at least one medical expert per language would be needed to judge the medical correctness.

Quality improvement *How can we further improve the quality of the presented corpus?*
First, although the pseudo-tweets are not the most natural regarding language-specific syntax, vocabulary, and how people generally write on social media, we observe some cross-cultural differences, also when compared to real tweets. For example, tweets in English, German, and French are usually written in a different "tone" than, e.g., tweets in Japanese.

It might not make a big difference in translating, for example, clinical guidelines from Japanese to English, since these follow a specific procedure and are more standardized and formal. In contrast, translating pseudo-personal messages needs further refinement to make them more natural in their respective language. Using one generative model per language would defy the goal of producing a quasi-parallel corpus. A possible option would therefore be to train a multi-lingual model on both translation and generation, which, given a drug name as seed, generates one tweet per language. Adding, for example, an additional loss function that penalizes the messages in the different languages diverging too far from each other with respect to their content might help keep the data parallel.

On a related note, it is also necessary to check how much of the real data is still preserved in the generated tweets, i.e., if they still contain any private information. This could be done, for example, with sentence embeddings (Reimers and Gurevych, 2019), as in Hiebel et al. (2023b).

Finally, finding outliers again, e.g., with respect to medical incorrectness or authenticity, and manually re-formulating them would create a much higher quality. This, of course, is very labor-intensive, but maybe newly emerged models like `ChatGPT` could be prompted in a way that allows to receive correct and more natural sounding tweets.

Usability *How useful is this corpus for (i) multi-lingual research in biomedical NLP and (ii) as a means to circumvent privacy issues with social media data?*

[51] E.g., "Is this tweet fluent in French? Decide on either yes or no, based on the following characteristics ... "

We have not yet conducted experiments using the described data as training material to allow predictions on other real datasets. However, regarding (i), it is to be expected that at least augmenting other corpora with our multi-lingual corpus should improve performance, especially on the binary detection of ADRs[52], since non-English data for ADR detection is scarce and imbalanced. Also, to the best of our knowledge, except for the KEEPHA dataset, which contains different data sources, there is no multi-lingual quasi-parallel corpus of that kind, allowing the community to advance multi-lingual research in the biomedical domain further. With respect to (ii), the corpus might be used as an exemplary dataset for training and evaluating models without accessing any real user data. This is especially useful in light of recent developments of large language models like `ChatGPT`: Since the generated tweets are not connected to an actual person, they can serve as intermediate training or fine-tuning material without the risk of exposing PHI to closed-sourced models.

Note that the presented data are limited because they are based on specific medication and disease names, as is often the case with Twitter data.

4.3 Summary & Conclusion

In this chapter, we discussed three data-centric topics. First, we demonstrated the development of the French and German part of a new tri-lingual corpus for detecting ADRs. We provide a 10,000-document-strong German dataset with binary annotations and a smaller French corpus, containing currently 864 documents translated from German. Both contain documents describing ADRs and are written from the perspective of laypeople.

The documents with a positive label are further annotated in more detail. These annotations include entity mentions, attributes, and relationships between these mentions. We do not only annotate ADRs, but provide entity, attribute, and relation types for everything relevant to the person describing their health issues. This includes, for example, time expressions to gain information about, e.g., the duration of a disorder, or information about the medication route. Also, we consider a patient's opinion or assessment about treatments, doctors, and their sense of well-being, allowing a unique, patient-centered view of the descriptions. ADRs in particular are represented by associating relevant entities (`drug`, `disorder`) with a `caused` relation, allowing also other medical signs or symptoms to be annotated and linked to relevant information, such as routes or mentions of body parts. The annotation guidelines and scheme were developed and tested in four languages: German, French, Japanese, and English. They are expected to be readily applicable to other languages and text collections since the already tested languages are quite different and we used different types of user-generated texts for their development. They therefore provide a good starting point for annotation of any health-related user-generated text and are meant to offer a common basis for more annotations, also for other teams. Summarized, we provide the first multi-lingual corpus for pharmacovigilance using user-generated text. This thesis presented the corpus at the time of writing, but note that more documents and possibly more annotations are to come, for example, the normalization of disorder mentions to medical ontologies.

In the second part of this chapter, we investigated the possibility of directly asking users online about their experience with ADRs in a prototype study. With this, we wanted to make sure that patients are granted the right to consent or not consent to the collection and distribution of their data, and establish whether people would respond at all. The study was based on a survey that we distributed on several social media channels and survey platforms, in both English and German. In fact, over the course of one month, we received about 30 usable responses, containing detailed descriptions of ADRs, medications, and diagnoses. Although this

[52] The multiple labels of each sample can be easily converted to a binary label.

is insufficient to compile an entire corpus, the responses might still be helpful to fine-tune or prompt large language models.

Finally, we analyzed the parallel data created for the NTCIR'17 Adverse Drug Reaction shared task. Since these data were generated in Japanese and translated into German, French, and English, we were interested to see if this corpus could substitute real data collected from social media. We found several issues highlighting the problems introduced by the generation, but mainly by the translation process. These issues stretch from unnatural texts over medical incorrectness to incomprehensibility and possibly invalid labels. But not all is lost with these data: Although oftentimes not comparable with authentic tweets, they can still be used for, among other things, data augmentation and few-shot experiments. Further, improving them partly manually might be worth the effort and would provide another dataset for multi-lingual detection of ADRs.

In conclusion, we investigated three ways of creating multi-lingual data for the detection of ADRs, all associated with different amounts of effort and quality. In combination, they provide a decent body of new data which should help to improve the work in cross-lingual and therefore cross-country pharmacovigilance using methods of Natural Language Processing.

Chapter 5

Document Classification

In this chapter, we describe experiments using the first part of the German dataset we created, LIFELINE-DE-1. The goal of the experiments was to leverage Transformer-based models to classify the given documents into the classes *positive* and *negative*, i.e., containing ADRs versus not containing ADRs. Since the German dataset was rather small at the time of this work, English data were used to help in the classification. Thus, in this chapter, we address RQ 3 and RQ 5.[1]

5.1 Datasets

The first dataset we use is the German corpus LIFELINE-DE-1, see Table 4.3. It is annotated with binary labels: The *positive* documents contain reported ADRs, while the documents categorized as *negative* do not contain any mention of ADRs. See Section 4.1.2 for the detailed annotation process. The major challenges of this dataset are the high imbalance of labels and the low number of samples overall. The positive class only comprises 101 documents, while the negative class is 4,068 documents strong, resulting in a positive-negative ratio of 1:40. Moreover, the topic distribution is imbalanced as well. Posts concerning women's health dominate, while there are less than 100 posts for topics such as nerves, nutrition, and men's health. See Figure 5.1 for the distribution of documents over topics and labels.

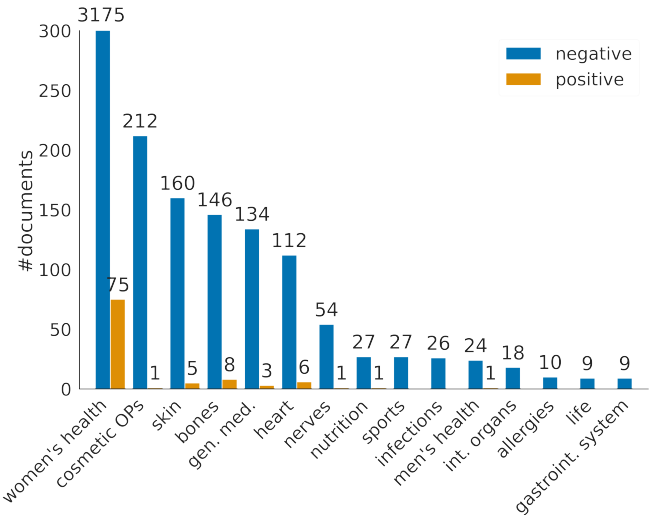


Figure 5.1: Distribution of documents over topics and labels in LIFELINE-DE-1. The y-axis is cut off to a low for a better visualization.

[1]This work was published in **Raithel** et al. (2022).

The second dataset is a combination of the English datasets CADEC (Karimi et al., 2015) and PSYTAR (Zolnoori et al., 2019), both based on the patient forum AskAPatient and created with a focus on ADRs. Together, they contain 2,173 documents of which 1,683 are positive and 454 are negative. Note that the label distributions are reversed compared to the labels of LIFELINE-DE-1. For a more detailed description of the English data, the reader is referred to Section 3.4. As a baseline, we use a "traditional" machine learning approach and compare the results with those of one English and one multi-lingual Transformer-based model fine-tuned in different settings.

Pre-processing Before feeding the documents into the models, they undergo a simple pre-processing. If still present, user names, URLs, dates, etc., are replaced by placeholders, e.g., `<URL>`. Then, the documents are tokenized either by white space (for the baseline models) or using the word-piece tokenizer (Wu et al., 2016) of the respective Transformer model. Based on preliminary experiments, the data is filtered for documents longer than four tokens and shorter than 300 tokens.

5.2 Methods

In the following, the tested strategies for document classification are described. As a baseline, we use a "traditional" machine learning approach and compare the results with those of an English and a multi-lingual Transformer-based model fine-tuned in different settings.

5.2.1 Baseline

The baseline is an SVM (Boser et al., 1992) classifier trained on the target data. We use an average of the word vectors created with the `fasttext` library (Bojanowski et al., 2017) to represent the input documents. Further, the class weights of the data are calculated to train the SVMs in "balanced" mode[2], otherwise default (hyper-) parameters are used.

5.2.2 Two-Stage Fine-Tuning

Since the presented dataset is small and imbalanced, and at the time of the experiments, there was no domain-specific Transformer model trained on the German language, we decided on a two-step approach for classification. First, we fine-tune the respective model on English genre-specific data (source language) and then add an optional second fine-tuning on German data (target language). The second fine-tuning is then further divided into three strategies, which we call `per_class`, `add_neg`, and `add_source` (described below).

Before fine-tuning, however, a Transformer-based model needs to be chosen. The perfect base model for this use case would have been one pre-trained on multi-lingual, user-generated, and health-related texts. Since this combination is not available (yet), we experiment with `BioRedditBERT` (Basaldella and Collier, 2019), henceforth `BRB`, a model trained on English user posts from specific health-related sub-Reddits, and `XLM-RoBERTa` (Conneau et al., 2020), henceforth `XLM-R`, a multi-lingual model trained on crawled text from the general domain, both equipped with a classification head. See Section 3.1.2 and Section 3.2.2 for a more detailed description of `XLM-R` and `BRB`, respectively. `XLM-R` was shown to work better than `mBERT` on several cross- and multi-lingual tasks (Hu et al., 2020; Adelani et al., 2021).

After conducting a hyper-parameter search optimizing for macro average F_1 score, the determined hyper-parameters and ten random seeds are used to fine-tune ten `XLM-RoBERTa` and

[2]This mode uses the labels to automatically adjust weights inversely proportional to class frequencies in the input data.

ten BRB models on the source language: XLM-R$_1$ – XLM-R$_{10}$ and BRB$_1$ – BRB$_{10}$. Using a number of different initialization seeds mitigates the instability of LMs as described in the work of Devlin et al. (2019) since being "lucky" with initialization might have a big impact on the results. We call these models *source language models* since they were fine-tuned on the English source data.

We assume that the high number of negative examples in the LIFELINE-DE-1 data will influence the model in favor of the negative class, and thus we evaluate several few-shot settings. In these, we mimic a "true" few-shot scenario (Perez et al., 2021) in that we also limit the number of examples for the development set. Therefore, our development set always contains the exact same number of examples per class as the training set. The following data combinations are explored:

per_class We balance the number of samples per class in the training and development set. This means that if we use, e.g., ten shots, the model is fine-tuned on five positive and five negative examples, and evaluated on five (different) positive and five negative examples.

add_neg In this scenario, we take n examples of the positive class (with n being the number of shots) and add a fixed amount of negative examples to the training and development set. For example, when using ten positive examples, we add $\{100, 200, 300, 400\}$ negative examples to the respective sets.

add_source Similar to the add_neg scenario, we add $\{100, 200, 300, 400\}$ negative examples and, in addition, $\{100, 200, 300, 400\}$ random examples from the source data to both training and development set. Adding source data might help to counteract the problem of *catastrophic forgetting* (McCloskey and Cohen, 1989) often observed in language models.

full All available (training) target data is used for fine-tuning to compare the performance of the models with those trained on the few-shot datasets. The resulting model is called XLM-R$_{full}$.

At least 20 examples seem necessary to conduct a reasonable evaluation, which leaves us in practice with 21. Thus, 80 examples are left for the remaining sets and thus, we can only use up to 40 positive shots for the described scenarios following the approach of a "true" few-shot scenario. Further, to reduce the number of experiments, we pick $n = 10$ and $n = 40$ positive examples for the implementation. Next, we create five different training and development sets by sampling with five different seeds from the described target training and development set. The test set stays the same throughout all experiments. With this data configuration, the experiments are carried out as follows:

1. The layers of the source language models, except their classifier, are frozen, and the classifier is fine-tuned again on the five sampled training sets within the few-shot settings described above. This results in the models XLM-R$_{fine_1}$ – XLM-R$_{fine_10}$ and BRB$_{fine_1}$ – BRB$_{fine_10}$.
2. For each scenario, each model is applied to the fixed target test set, and the final predicted classes (per scenario) are decided by majority vote.
3. The performance per scenario is averaged over the five seeds.

The experimental setup is visualized in Appendix C.3. As a last comparison, we apply the source language models in a zero-shot fashion to the target test set to investigate the impact of fine-tuning on the source language data versus no second fine-tuning at all.

Finally, we add some simple rule-based post-processing to the final predictions of the models by using an extensive German medication list and a self-compiled list of frequent abbreviations and keywords related to women's health. The medication list is a copy-pasted collection of 22,827 medication names from a German information website about health topics[3]. After determining the majority classification votes of the respective setting's models, each document's predicted class is checked. In case the class is *positive* but the document does not contain any of the collected medication names, the document's class is switched to *negative* Independently, we also switch the class to *negative* if the document is *positive* and contains a keyword from the women's health list since a lot of symptoms related to menopause can be confused with ADRs. The final scores are calculated for each approach. We are mostly interested in precision, recall and F_1 score of the *positive* class, but report both the negative and positive class scores and their macro average and Area Under the ROC Curve (AUC) score.

5.3 Results

The results of the above-described experiments are presented in the following.

5.3.1 Source Data (English)

The results for the first round of fine-tuning, i.e., fine-tuning only on English source data, are shown in Appendix C.1. We report final scores for each model and each seed (`XLM-R` in Table C.1, `BRB` in Table C.2) as well as their mean and standard deviation across seeds. Both models have a tendency towards the majority class, which is, in the source language data, the positive class, meaning that the documents containing ADRs can be detected by `XLM-R` with an average F1 of 91.03 and by `BRB` with an average F1 of 91.4, with both models being very close to each other in their performance. Note the low standard deviation of the scores for the positive class compared to the negative one, especially with respect to recall.

5.3.2 Target Data (German)

The results on the target data for the *positive* are displayed in Table 5.1, the result for the *negative* class are shown in Table C.4. The first block reports the performance of the SVMs, followed by the zero-shot approach, which is followed by the different settings of both `XLM-R` and `BRB`. Note that we omit the models and settings that scored an F_1 score of 0.0.

Baselines Within the SVM models, the best result ($F_1 = 17.39$ for the positive class) is achieved when training on all available target language data. All other settings for the baseline system score much lower, however, we can see a high recall for the positive class for both per_class settings. The per_class scenario with 40 shots even reaches the best overall performance with respect to recall for the positive class and has a very low standard deviation compared to the performance of the Transformer-based models.

Zero-shot The zero-shot models perform almost equally badly according to the F_1 score of the positive class. It is striking, however, that the multi-lingual `XLM-R` model achieves a much higher recall for the positive class (95.32), in fact, the second-highest recall over all experiments. The English `BRB` model, in contrast, gets the lowest recall for the positive class overall, showing a higher bias towards the negative class.

[3] https://www.apotheken-umschau.de/medikamente/arzneimittellisten/medikamente_a.html

model	method	target data	*positive class* P	R	F_1	*macro average* P	R	F_1	AUC
SVM	full	all	10.26	57.14	17.39	54.49	72.03	54.92	72.03
SVM	per_class	10	3.39 ± 0.65	85.71 ± 24.28	6.51 ± 1.25	51.36 ± 0.7	60.62 ± 7.64	27.99 ± 11.88	60.62 ± 7.64
SVM	per_class	40	3.23 ± 0.17	**99.05** ± **2.13**	6.26 ± 0.31	51.57 ± 0.14	60.64 ± 2.23	21.22 ± 3.48	60.64 ± 2.23
SVM	add_neg	10 + 200 neg	7.94 ± 1.16	40.95 ± 7.97	13.16 ± 1.3	53.11 ± 0.56	64.1 ± 2.64	52.78 ± 1.38	64.1 ± 2.64
SVM	add_neg	40 + 400 neg	6.16 ± 0.28	71.43 ± 10.1	11.33 ± 0.59	52.57 ± 0.3	71.53 ± 3.68	47.21 ± 0.68	71.53 ± 3.68
BRB	zero-shot	-	11.11	4.76	6.67	54.33	51.88	52.47	51.88
XLM-R	zero-shot	-	5.18	95.23	9.82	52.48	74.83	40.13	74.83
XLM-R	full	all	**57.64** ± **7.14**	28.57 ± 7.53	**37.52** ± **6.65**	**77.9** ± **3.55**	64.00 ± 3.68	**68.15** ± **3.33**	64.00 ± 3.68
XLM-R	per_class	10	5.24 ± 1.25	75.24 ± 31.66	9.75 ± 2.55	52.2 ± 1.08	70.84 ± 10.88	44.48 ± 1.9	70.84 ± 10.88
XLM-R	per_class	40	6.04 ± 0.87	93.33 ± 2.61	11.33 ± 1.55	52.87 ± 0.48	77.34 ± 3.65	43.58 ± 3.11	77.34 ± 3.65
XLM-R	add_neg	40 + 100 neg	26.37 ± 15.67	32.38 ± 30.38	19.81 ± 12.47	62.29 ± 7.67	63.44 ± 11.33	57.97 ± 6.94	63.44 ± 11.33
XLM-R	add_source	10 + 100 neg + 200 source	8.29 ± 0.8	81.9 ± 8.52	15.03 ± 1.26	53.84 ± 0.36	**78.95** ± **2.67**	50.55 ± 1.99	**78.95** ± **2.67**
XLM-R	add_source	40 + 300 neg + 300 source	15.84 ± 6.53	54.29 ± 18.01	22.55 ± 3.42	57.28 ± 3.1	72.6 ± 6.7	58.57 ± 2.8	72.6 ± 6.7
BRB	per_class	10	4.38 ± 0.7	41.9 ± 5.22	7.91 ± 1.12	51.21 ± 0.39	58.72 ± 1.98	46.58 ± 2.14	58.72 ± 1.98
BRB	per_class	40	4.74 ± 0.62	80.95 ± 6.73	8.94 ± 1.11	51.94 ± 0.35	68.77 ± 3.35	40.33 ± 4.2	68.77 ± 3.35
BRB	add_neg	40 + 100 neg	21.91 ± 20.74	9.52 ± 8.69	11.00 ± 8.6	59.79 ± 10.38	54.26 ± 3.86	54.66 ± 4.12	54.26 ± 3.86
BRB	add_source	40 + 100 neg + 200 source	24.21 ± 5.6	20.95 ± 4.26	22.23 ± 4.06	61.07 ± 2.83	59.59 ± 2.06	60.16 ± 2.08	59.59 ± 2.06

Table 5.1: Target language (German): results of the best runs for every scenario and for the positive class. We excluded those that scored an F_1 of 0.0 for the positive class. BRB = BioRedditBERT, XLM-R = XLM-RoBERTa. **P** is precision, **R** is recall and $\mathbf{F_1}$ is F_1 score.

full The second fine-tuning using all target training data, achieves, despite the high label imbalance, the best result for the positive class overall: an F_1 score of 37.52. It further reaches the highest precision for the positive class, but also, interestingly, the highest F_1 score for the negative class. This means that the negative examples (since we could not have more positive examples) helped the model to better distinguish between the two classes and to recognize positive documents with a higher precision (but lower recall). Note, however, the high variety in model performance over different seeds, as expressed by the standard deviation.

per_class In the scenario where the model has either 10 or 40 shots to learn from, we can only see a small difference in the performance. This is somewhat surprising but might be explained by the very small number of examples in general. In this scenario, the maximum number of training examples is 40, including both negative and positive examples. Even with a balanced label distribution, this seems not to be enough for a meaningful result.

add_neg Adding between 100 and 400 examples to either 10 or 40 positive ones does not help to improve the performance of the models: most result in an F_1 score of 0.0, and only the combination of 40 shots and 100 added negatives results in predictions of the positive class. Here, `XLM-R` performs better than `BRB`, probably drawing from its "knowledge" of German as a multi-lingual model.

add_source The second best result with respect to F_1 score for the positive class is achieved by the combination of 40 positive shots, 300 negative target examples and 300 random source examples (F1 = 22.55) with `XLM-R`, closely followed by `BRB` fine-tuned on 40 positive shots, 100 negative target examples and 200 random source examples (F1 = 22.23). These two were the only cases where the above-described post-processing improved the results, e.g., in the case of `XLM-R`, the F_1 score of the positive class increased from 22.55 to 28.56 and the `BRB` results increased from 22.23 to 25.33.

5.4 Error Analysis

Taking the predictions of the best-performing model ($\texttt{XLM-R}_{full}$), we analyzed the errors made by the model to get a better understanding of its performance and the data. The target test set (German) contains 824 documents. Of those, $\texttt{XLM-R}_{full}$ predicted 8/21 positives and 796/803 negatives correctly. Therefore, only 20 documents overall were predicted incorrectly. However, most of these documents belong to the positive class (false negatives), of which the model did not even get 50% right.

One of the falsely predicted positive documents was cut off before the person described their ADR issues, so the model did not have the correct information to learn from. The remaining documents contain some spelling mistakes and unclear formulations, but the documents are still perfectly understandable, at least by humans. However, a few documents mention ADRs only very briefly or implicitly. Conversely, some of the false negatives are very clear about the issues the reporting person has, and even the expression "side effects" is mentioned. In one of those, the reactions are described quite positively (weight gain) which might confuse the model in case it is biased towards documents with a more negative sentiment.

The seven false positives have clearer indications of why they were misclassified. In some of the documents, the reporting person is talking about side effects they experienced *before* they started taking a new medication about which they now report. Further, we find examples of health issues that can be easily confused with ADRs or where the reaction came from *not* taking the drug. Unfortunately, with the small test set, we can only provide anecdotal evidence for

the described assumptions as to why the model failed in predicting the correct class. A larger test set would provide more helpful information and might also reveal groups of errors.

5.5 Discussion

Based on the presented results, the classification of documents containing ADRs is still a problem far from being solved. Even using an English genre-specific model on English data, the results are not yet good enough to be reliable, mostly because of the strong label imbalance. Although the results for the English positive class are very good, they are rather low for the negative (minority) class. Also, there is a high variance across models detectable, especially for the recall of the negative class. This might be evidence that there are not enough examples for this class for the model to learn from reliably.

An interesting conclusion from the above experiments is that *more, but imbalanced* data works better in the presented use case than balanced data. On the other side, this conclusion must be investigated further, since the performance of the models might also heavily depend on the selected shot examples (are they representative of other positive examples?) and the total number of examples the model is supposed to learn from.

Adding source language data to the target training set seems to be an interesting direction as well, achieving the second-best performance after the *full* model. Combining the full source language dataset with the full target language dataset might therefore improve the results. With this, however, comes the question of whether adding examples, incorporating another language, or both is more helpful.

The $\texttt{XLM-R}_{full}$ model, although achieving the best F_1 score overall for the positive class, has a quite low recall for the positive class (28.57), compared to all other settings. This makes the model less useful than the other, overall worse-performing models since it cannot even be used to pre-filter relevant documents that have no labels yet. In most cases, the multi-lingual `XLM-R` model performed better or equally well than the mono-lingual `BRB` model. However, even though `BRB` was not trained on any German data, its performance comes close to the one of `XLM-R` in the add_source scenarios.

Another point of interest is the high instability of the models. This is likely due to the low number of training examples, but even for the full model, recall in particular has a high variance. A careful selection of "good" examples might help mitigate this issue, but finding out which examples are useful for learning at which step in the fine-tuning process is another point of investigation.

There are also issues introduced from the data. For example, with the high number of posts about menopause, a lot of symptoms described in the documents are not related to medication intake. Counteracting this with the medication list helped a little, but the distinction between drug-induced symptoms and other symptoms needs to be improved as well. This might be done by a better medication list which adheres more to the colloquial language used by laypeople but also by adding annotations of relations between drugs and symptoms as additional helpers. The latter, however, might create a vicious cycle: If we have only access to random, not pre-filtered documents, on which ones should we spend time annotating relations? We would first need to find the relevant ones, taking up the task of classification again. Also, related work (Magge et al., 2021) suggests that the classification step is still needed before extracting entities.

5.6 Summary & Conclusion

In the experiments described above, we tested several zero- and few-shot strategies in combination with cross-lingual transfer learning to classify German documents into those containing ADRs versus those that do not contain ADRs. We used an English Transformer-based model trained on health-related user posts from Reddit and compared it with a multi-lingual Transformer-based model trained on general domain data. The multi-lingual model outperformed the English model in almost all cases but is still not working well enough to be used as a reliable classifier.

Although many related works show good results when transferring knowledge from one language to another using multi-lingual models, this is not the case for the presented corpus. With respect to the transfer between languages (RQ 5), we can see some improvement when adding English source data to the German target data. However, the models still perform poorly overall. Of course, this does not only depend on the data but also on the models used. In our case, a multi-lingual domain-specific model, which does not exist yet, would most likely have performed better. Note that in the presented work, many challenges come together: a small dataset, a cross-lingual approach, non-availability of specific models, high label imbalance, ambiguous documents, and user-generated texts.

Regarding RQ 3, we cannot clearly answer the question. Although the full model (fine-tuned on all data) performs better than the rest in terms of F_1 score (for the positive class), there seems to be some benefit when combining shots with a certain amount of source and target data, especially in the direction of a better recall for the positive class, which, in practice, might be of more use than a mediocre model with a low score for filtering relevant documents. A careful selection of shots out of the few we have might be more beneficial than giving the model very complicated examples.

However, the amount of data in general was not enough to reasonably fine-tune and, in particular, evaluate the model. Therefore, in future experiments, we hope to achieve better results with the now extended corpus of about 10,000 documents.

Chapter 6

Medical Entity Extraction

Drug detection in biomedical texts is the task of extracting all drug mentions from the given text. In Example 6.1, an example from the CMED (Mahajan et al., 2021) data set is displayed, highlighting three drug mentions to be extracted. Next to it, we show an example from the newly created KEEPHA dataset in Example 6.2.

(6.1) *As a result of this, I think it is reasonable for us in addition to having her on atenolol to stop the hydrochlorothiazide, put her on ramipril and a nitrate.*

(6.2) a. ***de:*** *"Als ich damals mal Opri probiert habe, stand ich völlig neben mir. So richtig benebelt. ... Wie durch Watte durch den Tag..."*

b. ***en:*** *"When I tried Opri at that time, I was completely beside myself. Really woozy. ... like through cotton throughout the day..."*

The text on Example 6.1 is a very simple case for the state-of-the-art LMs as described below in Section 6.1. However, since medical records are not only written in English but also in other languages, Section 6.2 presents preliminary experiments on the transfer of knowledge with respect to drug mentions between Spanish, French, German, and English. Finally, in Section 6.3, the perspective is changed to user-generated texts (UGTs), which also can contain medical entities as can be seen in Example 6.2. Using the KEEPHA dataset, we evaluate a first baseline on these data, first only on drug mentions, and then for all annotated medical entities.

6.1 n2c2 Shared Task 2022

This section describes our participation in the n2c2 shared task 2022 track 1, *Contextualized Medication Event Extraction* on the Contextualized Medication Event Dataset (CMED) dataset[1] (Mahajan et al., 2021).
Disclaimer: I was mainly responsible for subtask 1, Medication Extraction.

6.1.1 Dataset

The dataset was built by Mahajan et al. (2021) to help the automatic understanding of *medication events* in clinical records. This is important to represent the patient's clinical history in context accurately. CMED contains 500 clinical notes overall, with 9,013 medication annotations[2] in total. Note that the underlying data was used before as the 2014 i2b2 / UTHealth Natural Language Processing shared task corpus (Kumar et al., 2015; Stubbs et al., 2015a,b).

[1] https://n2c2.dbmi.hms.harvard.edu/2022-track-1

[2] Note that this does not correspond to our counts, we found 7,230 (train/dev) and 1,764 (test) mentions, resulting in 8,993 mentions in total.

The new annotations[3] contain *contextual* annotations, i.e., the mentions of medications are further enriched by information whether a medication change is discussed or not (`Disposition`, `NoDisposition`, `Undetermined`), and if yes, in what context (`Action`, `Negation`, `Temporality`, `Certainty`, `Actor`). Since this section is only about medication extraction, the reader is referred to the work of Mahajan et al. (2021) for more details on the annotation.

Pre- & Post-processing

The data were provided in `BRAT` format and therefore, they first needed to be converted to a format suitable to the Huggingface transformers library (Wolf et al., 2020). The standard choice in this case is the `BIO` annotation. For this, we used the scripts provided by the `BRAT` maintainers[4]. Further, since the documents provided were rather long and did not fit into the 512-token limit of the employed models, they were split into chunks of sentences. The number of sentences per chunk and other hyper-parameters were determined using hyper-parameter search provided by the Weights & Biases framework (Biewald, 2020).

After pre-processing, the data was fed to the models. Since Transformer-based models are trained on sub-tokens, the pre-tokenized sequences were further split into sub-tokens, meaning that the token labels had to be aligned with the sub-tokens.[5] This process, chunking, tokenization, and sub-tokenization, had to be reversed in post-processing.

6.1.2 Models & Fine-Tuning

We considered several English pre-trained transformer models. However, after initial experiments, we chose `BioBERT` and `PubMedBERT` as our models to continue fine-tuning with[6] since they showed the best performance in terms of F_1 score on the provided development set.

All models were finally fine-tuned using early stopping, `AdamW` for optimization (Kingma and Ba, 2014), and five different seeds to account for training instabilities. The final submissions were fine-tuned on both the training and development set. We uploaded the following model combinations:

`BioBERT` ensemble: An ensemble of five `BioBERT` models with the same configuration but different seeds for initialization. All models used a chunk size of 30 sentences and a batch size of 8. A decision on a specific tag was reached via majority voting.

`BioBERT` combined with string matching: The BioBERT model with the best performance on the development set combined with a very simple string matching approach to find missing medication names that were longer than five characters[7]. String matching consisted of a string match between all drugs collected in the training and development set against the test set.

`PubMedBERT` ensemble: An ensemble of the best two `PubMedBERT` models trained on two different configurations, the main differences being the number of sentences per chunk, the batch size and five seeds each.

Contrary to the very good performance of the `BioBERT` models on the development set, the `PubMedBERT` ensemble achieved superior performance on the test set released by the shared

[3] 120 out of 500 documents were double annotated, however, Mahajan et al. (2021) only report the IAA for the context information.

[4] `BRAT` to `BIO`: https://github.com/spyysalo/standoff2conll, `BIO` to `BRAT`: https://github.com/nlplab/brat/blob/master/tools/BIOtoStandoff.py

[5] Each sub-token of a token received the same label as the entire token, i.e., all sub-tokens had the same label.

[6] The exact model versions we used can be found in Table A.1.

[7] The string length was determined experimentally.

task organizers and was our best submission with a strict F1 score of 0.9474 and a lenient F1 score of 0.9704. These results achieved the tenth place out of 32 participating teams overall, with all teams achieving scores very close to each other[8].

PubMedBERT	*strict*			*lenient*		
	precision	**recall**	**F_1**	**precision**	**recall**	**F_1**
model 1	0.9444	0.9444	0.9444	0.9660	0.9660	0.9660
model 2	0.9499	0.9342	0.9420	0.9729	0.9569	0.9648
ensemble	0.9407	0.9541	0.9474	0.9637	0.9773	0.9704

Table 6.1: The results of the best model: the ensemble of the best two PubMedBERT models.

Looking at only the scores, which are rather high, gives the impression that extracting the drugs out of the given texts works well, and ensembling the described models improved the results even a bit more, especially with respect to recall. However, during the error analysis, we found some interesting mistakes made by the models during inference, which are described below.

6.1.3 Error Analysis

In total, there were 45 (strict: 57) false positives, i.e., expressions wrongly labeled as drugs, and 33 (strict: 40) false negatives, i.e., drug mentions not recognized as drugs. Note that 12 out of 57 false positives and 7 out of 40 false negatives were due to boundary issues. In the following, examples of the found errors are given, highlighting the problems the best model experienced during prediction. First, examples of false positives are listed and categorized in different error groups.

Annotation inconsistencies 7 out of 45 false positives were due to annotation inconsistencies, e.g., “Phenytoin”, “hypertonic saline”, or “pulmozyme”, which were not annotated, i.e., those mentions are actually not wrong.

Spelling mistakes Words that became more “complicated” due to spelling errors were extracted as drug mentions, e.g., “probabyl” instead of “probably”, “takien” instead of “taken”. This also happened for non-standard English doctors’ names and might be a clue for the models not really “understanding” the context in which the medication mentions occur but rather learning features of the drug names themselves. For instance, the doctors’ names usually were written at the very end of the report, in a context where drugs were not mentioned anymore. Note, however, that, for example, “byl” was not a common ending of drugs in the training data.

Medical terms which are not drugs Some false positives might have been based on context, discounting the assumption above: Medical expressions, which were not drugs but used in a similar context, were extracted as well, e.g., “Tegaderm” or “ace”. This was also the case for mentions containing the expression “anti”, as in “anti-coagulated”.

The false negatives can be grouped into the following categories:

[8]The best team won with a strict F1 score of 0.9716 and a lenient F1 score of 0.9846.

Missed medication names Almost half of the false negatives (16 out of 33) were missed medication names, e.g., "Lisinopril", "MOM". We do not have an explanation as to why exactly those were missed when similar occurrences were detected. It might, however, also interrelate with the pre- or post-processing of the data.

Medical treatments 6 out of 33 false negatives were medical treatments like "chemo" or dietary supplements like "K" (probably for "potassium") or "Mg"(probably "magnesium", but also occurs as "milligram"). Admittedly, these are very difficult to detect correctly, especially the abbreviations of chemicals.

Spelling mistakes Spelling errors were another source of errors that led to some drug mentions not being recognized (e.g., "SSIR" instead of "SSRI"), giving evidence that the model might have relied on features coming from the medication names themselves.

Abbreviations Short versions of medication names, for instance "tobra" instead of (probably) "Tobramycin", were also missed by the model.

Ambiguous mentions Another common mistake observed were mentions that could be both body substances but also medication, e.g., "insulin" or "oxygen". There, again, the context plays a crucial role.

Some mentions also fell into both false positives and false negatives. For instance, mentions containing the term "medication(s)" (e.g. "hormone medication" (false positive) or "pulmonary medications" (false negative)) were sometimes annotated and sometimes not. Similarly, some medications or treatments occurred both as false positives and false negatives, for example, "nebs" and "methadone".

6.1.4 Summary & Conclusion

In conclusion, we found that a very simple but domain-specific Transformer-based model without any tweaks already worked very well on the provided data. However, the simple string matching we added in post-processing was too coarse and only hurt performance. Furthermore, combining the models trained with different seeds for initialization improved overall performance as well since this mitigated the mis-predictions of single models. Based on the error analysis, we found that the system identified more entities wrongly as drugs than it missed true entities. For those that were missed, we highlighted some examples, showing that these were often more difficult mentions like abbreviations and medical treatments. However, half of the false negatives were still missed, and for those, we cannot provide an explanation since most of these were recognized in other notes.

When working on these data, it became clear that they are homogeneous, i.e., one clinical note is very similar to another within this dataset. Further, as Mahajan et al. (2021) note, the data are not representative since they only contain clinical notes about patients with diabetes and heart diseases, leading to the same medication groups occurring in the texts. They are, moreover, based on only one data warehouse[9] (Mahajan et al., 2021), leading to the same structure of every note.

This homogeneous structure might have been learned by the models trained on the clinical notes. This will probably result in a degraded performance when applied to datasets from different providers or hospitals. Furthermore, the dataset is only available in English.

[9] https://www.medicalrecords.com/mrcbase/emr/partners-healthcare-longitudinal-medical-record-lmr-511-partners-healthcare-system-inc-4302007

The limitations of the presented work, however, gave rise to a new question: What is the performance of "simple" models as the ones presented on other languages, as well as across these languages? This is discussed in the next section.

6.2 Cross-lingual Drug Detection

As demonstrated in the section before, BERT-based models seem to work very well on clinical records such as the CMED dataset. However, a model working well on English data does not necessarily mean it works well on data in other languages. This section, thus, discusses the ability of *multi-lingual* Transformer models to improve medication detection in languages other than English. In more detail, the experiments conducted serve to gain first insights into how well the mentioned models, in combination with the selected datasets, work in general for the given task, i.e., can a multi-lingual model reliably detect drug mentions in texts coming from different sources and based on different annotation guidelines (RQ 4)? Furthermore, the experiments are designed to understand how the different languages, in this case represented by datasets, contribute to medication detection in another language.

6.2.1 Datasets

The datasets were selected based on availability and annotation. They were required to provide annotations on the entity level, describing medication names or similar, closely related types, such as substances. However, there are only a few available medical datasets in languages other than English to choose from, and in the end, we chose corpora in two Germanic (English and German) and two Romance (French and Spanish) languages.

All corpora are described in the following; the entity labels relevant to our experiments are boldfaced. They contain medication names (and sometimes chemicals) used in clinical texts, e.g., patient records. Usually, there is only one label per dataset dedicated to the desired expressions; sometimes, however, these labels cover a broader scope than only drug names.

German

BRONCO150 (Kittner et al., 2021) The Berlin-Tübingen Oncology Corpus[10] contains 150 discharge summaries of cancer patients who received treatment at either Charité Berlin or Universitätsklinikum Tübingen in Germany. The summaries were manually anonymized, split into sentences, and scrambled to avoid the possibility of tracing back discharge reports to individuals. The sentences in this corpus are annotated with three entity labels ("diagnosis", "treatment" and "**medication**") and normalized to terminologies. Only complete tokens were annotated, even if a sub-token was part of a medical entity. The authors define a medication as "a pharmaceutical substance or a drug that can be related to the Anatomical Therapeutic Chemical Classification System (ATC[11])" (Kittner et al., 2021).

GERNERMED (Frei and Kramer, 2022) This corpus[12] originates from the n2c2 2018 ADE dataset (Henry et al., 2020) which consists of annotated English EHRs covering several clinical entities such as "**Drug**", "Dosage", "Strength" etc. The German data samples are obtained through automatic machine translation, while annotation information is transferred into German using word alignment estimation. Therefore, it is not a gold standard dataset. In this

[10] https://www2.informatik.hu-berlin.de/~leser/bronco/index.html
[11] www.dimdi.de/dynamic/de/arzneimittel/atc-klassifikation/
[12] https://github.com/frankkramer-lab/GERNERMED

work, we use an updated dataset iteration available from the authors. According to the respective n2c2 annotation guideline, the drug entity should include all kinds of drugs except for "illicit" drugs and alcohol.

GGPONC v2.0 (Borchert et al., 2022) This datatset[13] is a data collection based on clinical practice guidelines in German. GGPONC is a collection of curated scientific text documents, i.e. clinical guidelines that include, for example, instructions for treating breast or lung cancer. It does not contain any personal data and thus is freely accessible. The entities labeled in this corpus are "Finding", "**Substance**" or "Procedure". The "Substance" label includes "general substances, the chemical constituents of pharmaceutical/biological products, body substances, dietary substances, and diagnostic substances (...)"[14].

Ex4CDS (Roller et al., 2022) The dataset[15] consists of short notes written by physicians in the context of estimating patient risks. The text data has similarities to clinical text, was annotated with entities and relations, and comprises entities like "Condition", "Lab Values", "HealthState", "Measure", or "**Medication**". Medications are defined as generic drug names, groups of medications, and active substances.

English

CMED (Mahajan et al., 2021) CMED[16] was published by the organizers of the n2c2 challenge in 2022. It contains over 500 clinical notes based on the 2014 i2b2/ UTHealth NLP shared task corpus (Stubbs et al., 2015a,b; Kumar et al., 2015) and is annotated with medication changes. It is already described in Section 6.1.1.

French

Quaero (Névéol et al., 2014) The Quaero French Medical Corpus[17] was designed for medical named entity recognition and normalization in Medline titles and EMEA documents. The types of clinical entities follow the UMLS semantic groups and allow labels such as "Anatomy", "**Chemical**", or "Disorder". The label "CHEM" contains chemicals and drugs as defined by Bodenreider (2004), including, for instance, antibiotics, clinical drugs, elements or enzymes, amongst others.

DEFT (Grouin et al., 2019) The DEFT corpus[18] contains more than 700 documents from freely available clinical case reports in French and is a subset of the CAS corpus (Grabar et al., 2018). The data are classified into four general categories ("age", "gender", "outcome" and "origin"). A subset of the reports is then annotated in a more fine-grained way, using, for instance, entity labels relating to physiology (e.g., "body measurement") or surgeries (e.g., "surgical approach" or "medical device"). The entity we are interested in is the one named "**substance**", a subset of the broader category of drug annotations, including labels like "concentration" or "mode". "Substance" is defined as "commercial and generic drug names or generic substance" (Grouin et al., 2019).

[13] https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/

[14] Annotation guidelines of GGPONC: https://github.com/hpi-dhc/ggponc_annotation/blob/master/annotation_guide/anno_guide.pdf

[15] https://github.com/DFKI-NLP/Ex4CDS

[16] To the best of our knowledge, these data are not (yet) publicly accessible.

[17] https://quaerofrenchmed.limsi.fr/

[18] https://deft.limsi.fr/2019/index-en.html

Spanish

PharmaCoNER (Gonzalez-Agirre et al., 2019) This corpus[19], developed for the PharmaCoNER shared task, contains approximately 1,000 manually annotated clinical case studies in Spanish. The annotated entities are "**Normalizables**" (mentions of chemicals[20] that could be manually normalized to a CUI, "**No_Normalizables**" (mentions of chemicals that could not be normalized), "Proteinas" and "Unclear".

CT-EBM-SP (Campillos-Llanos et al., 2022) The Clinical Trials for Evidence-Based Medicine in Spanish corpus[21] is annotated with entities from UMLS. The texts are taken from journal abstracts about clinical trials (500 documents) and announcements of trial protocols (700 documents), containing entities belonging to categories such as "Anatomy", "Pathology", or "**Chemical**". The latter are defined as "pharmacological and chemical substances" (Campillos-Llanos et al., 2022).

In total, we consider four German, one English, two French, and two Spanish datasets. All these are based on similar but not identical annotation guidelines and annotate some kind of medication mention. Note that although the guidelines might be comparable, the data were created with different goals in mind, by different annotators and in different settings. Therefore, the scope of the annotated entities might vary or include or exclude particular expressions. An overview of the datasets and the number of annotated entities is shown in Table D.2. The number of entities labeled as some variety of `drug` are detailed per training, development and test set.

6.2.2 Methods

In this section, pre-processing and fine-tuning strategies are laid out.

Pre-processing

If no pre-defined dataset split was given, the corpora were divided into a training (70%), development (15%), and test set (15%) on document level where possible. To re-use the preprocessing pipeline described in Section 6.1.1, all data not yet in BRAT format were converted to that effect. Further, all medication-related labels described above are mapped to the label `drug`. Again copying the procedure in Section 6.1.1, the texts are split into sentences and aggregated into chunks to fit into the 512 sub-token limitation. Each chunk contains up to 26 sentences, which was determined experimentally.

Models

We again rely on `XLM-RoBERTa` in its base version. We compared it with the large version and found that in the case of drug detection, the final performance was very close for both models, but fine-tuning the larger model took longer. Every model in every setting is fine-tuned using five different seeds.

Since the datasets have different sizes and each contains a different number of annotated entities, we apply *weighted random sampling*[22] to the batch generation process. For this, we calculate the number of samples per language in the entire training set and use the inverse weights for sampling. This makes sure that the German samples are not preferred over the

[19] https://temu.bsc.es/pharmaconer/
[20] For this dataset, the terms "chemical" and "drug" are used interchangeably.
[21] http://www.lllf.uam.es/ESP/nlpmedterm_en
[22] https://pytorch.org/docs/stable/data.html#torch.utils.data.WeightedRandomSampler

samples of the other, smaller datasets when filling the batches, avoiding the models learning a bias towards one language or entity type. However, due to the small size of some datasets, it might happen that one sample of a particular language and dataset might be seen several times during fine-tuning. We did not apply this method based on the dataset but only based on language. Fine-tuning details are provided in Appendix D.1.3.

The test set predictions of the models based on the different seeds are ensembled and decided via majority voting. The final performance is evaluated using the n2c2 evaluation script, which returns strict and lenient scores for precision, recall, and F_1 score.

Experimental Setup

We run three different experiments to compare models fine-tuned on different data combinations.

Mono-lingual We fine-tune five *multi-lingual* models for each language separately. These models are then applied to *all* test sets, and the predictions are ensembled by majority vote. For instance, a model fine-tuned on only the German datasets is evaluated on the German, English, French, and Spanish test sets. Note that we do not take these models in the sense of a "lower bound" but only as a comparison. We could have also chosen real mono-lingual models but decided against them to rule out different pre-training strategies or datasets used by language-specific models. Presumably, the baselines as we chose them might even perform better than the cross-lingual or multi-lingual models since they are fine-tuned within language. However, we do not want to create a new state-of-the-art but investigate the *current* state-of-the-art to see in which directions to improve cross-lingual medication detection.

Multi-lingual For the multi-lingual experiments, `XLM-RoBERTa` is fine-tuned on all training sets in all languages, again using five different seeds. All medication labels are mapped to one common label `drug`. The predictions of all models on all test sets are again ensembled. These models show the current performance of multi-lingual models on different datasets in a specific domain.

Clusters Finally, we compare the above-mentioned setups with models fine-tuned on language clusters. These clusters are similar language groups, i.e., the Romance group (fr, es) and the Germanic group (de, en). The development data is divided into these clusters, while the final test data is the same as in the other setups. With this setup, we would like to investigate whether there is a benefit in using only similar languages, i.e. whether this allows the models to perform better in medication extraction.

6.2.3 Results

This section provides the results of the above-described experiments using exact and lenient precision, recall, and F_1 score. If not otherwise stated, we refer to the lenient scores since these are less sensitive to span boundary errors. The strict scores are shown in Appendix D.1.4. We refer to models fine-tuned on language x and evaluated on language y as $m_{x,y}$, the same for F_1 score ($F_{x,y}$).

Mono-lingual The results of the models trained on only one language are shown in Table 6.2. The table shows the performance on the datasets in the language the model was trained on as well as on the combined test corpus containing all languages (*all*). The results are further partitioned by language.

		lenient			
train	**test**	**precision**	**recall**	**F_1**	Δ
	all	73.3	81.3	77.1	+7.7
	de	85.6	87.4	**86.5**	0.0
de	en	68.7	87.0	76.8	-18.1
	fr	57.3	57.5	57.4	-7.1
	es	67.3	80.1	73.1	-16.4
	all	74.0	63.0	68.1	+16.8
	de	64.6	59.8	62.1	-24.4
en	en	96.3	93.4	**94.9**	0.0
	fr	61.0	41.4	49.3	-15.3
	es	78.5	59.0	67.4	-22.1
	all	75.2	64.4	69.4	+15.5
	de	75.6	63.5	69.1	-17.5
fr	en	75.2	67.8	71.3	-23.6
	fr	67.1	62.2	**64.5**	0.0
	es	79.1	64.5	71.1	-18.4
	all	79.2	72.5	75.7	+9.2
	de	75.7	68.8	72.1	-14.4
es	en	80.4	68.4	73.9	-20.9
	fr	63.2	55.4	59.1	-5.5
	es	90.1	88.9	**89.5**	0.0

Table 6.2: Results of models trained on the single languages. The evaluation scores are reported as micro average scores over all test set samples and separated by language. The best scores on a *test* language are marked in bold font. The rightmost column shows the difference (Δ) in F_1 score between the current model and the (next) best model on this test set. For example, "+7.7" indicates that the better model is 7.7 percentage points better than the current model, while "-18.1" indicates that the next best model is 18.1 percentage points worse than the current model. *all* refers to the combination of all test datasets.

The best-performing training language as evaluated on the test set is highlighted in boldface. Unsurprisingly, this is always the language the model was fine-tuned on. The distance between the best F_1 score and the second best F_1 score on the respective dataset is shown in the rightmost column. For example, $F_{de,de} = 86.5$, i.e., the model fine-tuned on German and tested on German ($m_{de,de}$). In comparison, the model that comes closest to that is the one fine-tuned on Spanish: $F_{es,de} = 72.1$, performing 14.4 points worse than $m_{de,de}$.

Note that for all languages except German, precision is always higher than recall. For the model trained on only German, recall is always higher than precision, no matter the test dataset. Finally, the model trained on German performs best (within this group) when tested on the corpus containing all datasets.

Multi-lingual Considering the result of the model fine-tuned on all data in Table 6.3, we find that its performance on all ($m_{all,all}$) is higher than when using only a single language for fine-tuning, which makes sense. Further, the scores per language-specific test set, e.g., those of $m_{all,fr}$, are slightly below those fine-tuned only on the language's respective training set, but often only by a small margin: 1 percentage point for German, 1.9 for English, 1.5 for French, and 1.1 for Spanish. For all languages except for French, precision is always lower than recall.

Note that for English, the recall increases when using all data for fine-tuning, while precision drops by 5.6 percentage points. For all other languages, both precision and recall decrease in this setting.

		lenient			
train	**test**	**precision**	**recall**	**F_1**	**Δ**
	all	83.8	86.0	**84.8**	0.0
	de	84.2	86.9	85.5	-1.0
all	en	90.7	95.4	93.0	-1.9
	fr	66.7	59.6	63.0	-1.5
	es	85.7	91.4	88.4	-1.1

Table 6.3: Results of the multi-lingual model trained and evaluated on all languages. The Δ column shows the distance in performance to the best model on this test set. The best F_1 score on the combination of all datasets across experiments is marked in bold.

The results are partitioned by dataset in Table D.6. Here, it becomes clear that the low performance on French mainly comes from the low performance on the DEFT dataset, mostly due to a very low precision (18.6%). For German, the lowest performance is on the Ex4CDS 2.0 dataset, mainly because of a low recall. The results on all other datasets seem decent, considering that the lowest score is achieved on Quaero with $F_1 = 71.6$ using the multi-lingual model fine-tuned on all languages.

Clusters In Table 6.4, the results when fine-tuning on language clusters are laid out. The highest scores within this experiment setting are bold-faced: When testing on all data, the combination of German and English seems to work best, but still shows a difference of 4.4 percentage points to the model fine-tuned on all languages (see Table 6.3). This combination also works better than the combination of French and Spanish. This is reversed for the French and Spanish test sets, where the subset of Romance languages achieves higher scores.

Note that the combination of German and English is better on German overall than when adding French and Spanish, i.e., the cluster model m_{de+en} achieves better results on the German test set than the model fine-tuned on everything ($F_{de+en,de} = 86.4$ versus $F_{all,de} = 85.5$). The model fine-tuned on German only is slightly better: $F_{de,de} = 86.5$. Table 6.4 also shows that fine-tuning on only French and Spanish results in a higher F_1 score for the French test set than fine-tuning on all languages. The only language that does not benefit more from the cluster-based approach than the multi-lingual approach is Spanish: Using the clusters, the performance is 2.4 percentage points worse while only 1.1 percentage points worse when using the model fine-tuned on all data.

Discussion

The results are now discussed in more detail and set into context with the data and model fine-tuning procedure.

Mono-lingual In the mono-lingual cases, the results are not surprising: The models trained on only one specific language perform best on exactly this language. In this set of experiments, we cannot find evidence that any other language comes even close to a similar performance. However, it would still have been possible that adding other languages, e.g., English data to German, would have improved the results for both English and German. This is not the case

		lenient			
train	**test**	**precision**	**recall**	**F_1**	Δ
de, en	all	81.7	79.1	**80.4**	-4.4
fr, es	all	74.3	77.3	75.8	-9.1
de, en	de	86.3	86.6	**86.4**	-0.1
de, en	en	92.6	94.8	**93.7**	-1.2
de, en	fr	60.3	52.2	56.0	-8.6
de, en	es	76.6	71.2	73.8	-15.7
fr, es	de	74.2	72.8	73.5	-13.0
fr, es	en	66.7	77.1	71.5	-23.3
fr, es	fr	65.3	62.6	**63.9**	-0.6
fr, es	es	83.3	91.3	**87.1**	-2.4

Table 6.4: The lenient results of the cluster approaches. The rightmost column shows the difference in the best score on the respective test set across all experiments.

and might hint that the datasets are too diverse to support each other. Remember the homogeneity of the English CMED data described in Section 6.1.1: It might be simply too different in structure and context to help in the detection of further German medication mentions.

For German, where recall is always higher than precision, no matter the test set language, this might be due to the high number of German entities provided in the training data: There are in total 26,849 medication mentions distributed over four corpora. This might allow the model to learn about many contexts in which the medication mention can occur, but also about "wrong" contexts, that is, an over-generalization. On the other hand, a low(er) precision means that many of the detected entities do not appear in the gold data, i.e., there are many false positives. The difference between precision and recall is particularly striking for the model evaluated on English and Spanish. See, for example, the results of $m_{de,en}$: The recall score is 18.3 percentage points higher than the precision score. This might mean that the model identified a lot of potential entity spans, and many were correct (recall is 87.0%). Still, at the same time, the model also predicted many spans that were not correct, and therefore, precision is only 68.7%. Therefore, we might assume that when using the large German dataset for fine-tuning, the model learns many potential positions of `drug` entities, but not all can be transferred to the other languages' datasets. To further investigate this, it would be necessary to identify which spans were predicted falsely as medication when the model is trained on German data only.

The resulting scores are reversed when considering the other languages: When fine-tuning on English, French, or Spanish, the model's ability to find potential `drug` candidates is reduced, but those it finds are – in some cases – correct.

Finally, the model fine-tuned on German data only achieves, within this set of experiments, the best score when applied to the corpus containing all data. This is mostly also due to the contribution of the German data in the training set (i.e., many different examples), as well as the fact that a big portion of the test set is German as well. It would therefore be better to balance the test set based on language and dataset.

Multi-lingual We find that when the model is fine-tuned on all languages, the performance on the entire dataset ("all") is the best when compared across experiment settings, i.e., fine-tuning on monolingual data or on clusters shows a lower performance. Note that we made sure that in every batch each language is at least represented once according to their inverse weights. Therefore, we cannot say that the German data, represented by four datasets and having the

strongest foundation in terms of entities, is responsible for this, but rather the combination of all datasets. Separated by language, the German datasets (except for Ex4CDS), the English one, and also the Spanish ones receive good scores overall (Table D.7).

Clusters The cluster approach was based on the assumption that closer languages might benefit more on the performance than languages that are more distant to each other. In the experiments, this was true for all languages but Spanish: Fine-tuning on English and German improved the performance on both the English and German test sets compared to the multilingual approach. The same is true for the cluster of Spanish and French, which improved the performance on French when compared to the model fine-tuned on everything. For Spanish, the model fine-tuned on all data is still better than the one fine-tuned on the Romance cluster, leading to the assumption that either the French data introduces errors or the English and German data provide some information that is also useful for the Spanish corpora. However, it might also be the reverse: Removing "distracting" datasets from the data configuration makes it easier for the model to adapt to the closer languages, achieving better results but probably reducing the robustness of the system.

We also assumed that this approach might improve over the mono-lingual models, adding a more diverse context and more examples in close languages. However, this did not happen, the mono-lingual models consistently achieved the best scores across all experiments.

Summary

We find that the mono-lingual models achieve the best performance for each individual language, but the model fine-tuned on data in all languages achieves very close scores. When dividing the datasets based on their language family, they achieve better results than the models fine-tuned on all languages except for Spanish.

Several variables need to be considered and controlled for: First of all, there are four different languages distributed over several datasets. Second, those datasets come from different sub-domains, i.e., discharge summaries (BRONCO150), clinical notes and EHRs (GERNERMED, Ex4CDS, CMED), guidelines (GGPONC v2.0), and scientific documents (Quaero, DEFT, PharmaCoNER, CT-EBM-SP). They further annotated slightly different entity types. Although these are all related, they might exclude or include mentions that are included or excluded in the other datasets, making it difficult for a model to learn consistently. Therefore, it is no surprise that the mono-lingual models perform well on each language separately: There is less distraction and they can quickly adapt to the specific language. This is also true for the cluster-based approach. However, this can come at the cost of a lower robustness, making it more difficult to apply the models reliably on other languages, which is our main interest. We therefore conduct an error analysis on the predictions of the multi-lingual models to investigate where the problems lie.

6.2.4 Error Analysis

Since we are interested in a model that is reliably applicable to several languages, we conducted an error analysis on the multi-lingual model that was fine-tuned on all available data. The error analysis is qualitative, focusing on the lenient predictions of the model. In Table 6.5 the numbers of false positives (FPs) and false negatives (FNs) are shown.

False Positives

The false positives are discussed first. These mentions were predicted as (part of) a drug name but are incorrect according to the respective dataset's gold annotation.

language	FPs	FNs
de	376	382
en	113	63
fr	298	175
es	287	142
total	1074	762
total (unique)	977	755

Table 6.5: False positives (FP) and false negatives (FN) for the multi-lingual model.

Annotation Errors Out of the collected false positive samples, several can be considered as *true* positives – however, not according to the ground truth of the underlying dataset. For example, on the DEFT dataset, the model predicted, among other things, "Rivotril" and "paroxétine", both of which are, indeed, names of medications. However, they were not evaluated as correct for the respective dataset.[23] Investigating the occurrences of the entities, we find that "Rivotril" only occurs in the Spanish training data and in no other dataset. "Paroxetine", however, can be found in the training data of GGPONC and GERNERMED ("Paroxetin"), CMED ("Paroxetine"), PharmaCoNER and CT-EBM-S ("paroxetina") and even in DEFT ("paroxétine"). Similar examples for German would be "Dopamin" (GGPONC) or "Metamizol" (BRONCO150, GGPONC); both were not labeled in the ground truth in some cases. However, we could verify them to be present in the training sets of GGPONC, PharmaCoNER, BRONCO150, and CT-EBM-SP. Consequently, we assume these to be annotation errors or entities that were not relevant for the respective corpus for some reason. Also, this speaks for the model since it recognized these entities, even if they were not "officially" correct. It also makes the model a good instrument for validating annotations since it can highlight those that are missing.

Groups of other medical terms In the FPs across all languages and datasets, we can find terms that belong to specific groups. These groups and their members often have medical associations but are not medications themselves, similar to what we saw in the error analysis in Section 6.1. However, their medical "context" might be a reason for their prediction. Some of the most visible groups, i.e., those where we find more than one representative per language, are shown in Table D.9. They contain proteins, chemical compounds, abbreviations, general medical expressions, medical terms and tools that are not drugs, and dietary supplements.

A reason for these predictions might be the label definitions of the different datasets. Some of them, e.g., Quaero and PharmaCoNER, include enzymes or chemical substances in their respective entity definitions. Also, the mentioned expressions are all used in very similar or even the same context as drugs, and therefore, the model might not be able to distinguish them semantically from medications.

Summary Summarizing the analysis of false positives, we observe that most of the incorrectly detected expressions can be categorized into a particular group. Most of these classes can be associated with medicine, medical treatments, or other things related to a clinical setting. Some FPs are simply based on annotation errors or on minor differences in the dataset guidelines (e.g., "CHEM" versus "Medication"). Only very few can not be explained at all and might be just a coincidence based on the context in which they occur. In those cases, it would be interesting to check the certainty of the model for its prediction.

[23] Note that some of the DEFT examples were only annotated partially.

False Negatives

Next, we consider medication mentions missed by the system.

Groups of medications and other medical terms Some FNs can be, similar to the FPs, categorized into several groups. We find examples of therapies, general medication expressions, brand names, mixtures of medication names and their routes, and ambiguous or very short mentions. See Table D.10 for some examples.

Finally, most FNs seem to be actual medication names (e.g., ***de:*** *"Avelumab"*, ***en:*** *"LISINOPRIL"*, ***fr:*** *"Atripla"*, ***es:*** *"folato"*) which were not detected by the system. The reason might be that these drugs were never seen in any training examples, or that the context in the test example did not match the one the system was trained on. Regarding the former, this is not the case for the examples provided above, only "Atripla" occurs merely twice in the training data, all the others are quite frequent, with "Lisinopril" being mentioned at least 188 times in both German and English data. "Brennnesselsaft" (de, ***en:*** *"nettle juice"*), by contrast, was indeed never seen during training, it only occurs in the test data. In fact, 440 of the 755 FNs were absent in the training data.

Nevertheless, in any case, a model should be capable of generalizing to unseen mentions, as long as the context in which they occur is similar. However, since we merged the datasets, context might exactly be the problem: `substance`s, for instance, might occur in medication contexts and other situations. Among the mentions not in the training data, we find expressions such as "lokale Strahlenträger" (de, ***en:*** *"local radiation carriers"*[24]), "Zigarettenrauch" (de, ***en:*** *"cigarette smoke"*) or "tabanidés" (es, ***en:*** *"tabanidae"*, some kind of fly).

This expression "Brennnesselsaft" is another good example of the differences in the annotation: It is certainly debatable whether "nettle juice" can be really seen as a medication, and indeed, in the GGPONC v2.0 corpus, it is annotated as `Substance`, which might be more correct than `drug`. As mentioned before, about 50% of the false negatives are from the German data, where GGPONC v2.0 represents the biggest part. In contrast to the other German datasets, GGPONC v2.0 also provides a broader scope for medication expressions, using `substance` for annotation.

General Observations

We conclude the error analysis with general observations on the predicted entities.

Span Length The system seems to have difficulties in deciding the span length of an entity. In terms of scores, this is ignored in lenient mode (since a match does not have the exact offsets as in the ground truth document), but some lenient true positives are conspicuously longer than they need to be from the perspective of annotating medication names. This might be due to the strikingly different span lengths across the training datasets: In GGPONC, PharmaCoNER, CT-EBM-SP, Quaero, and DEFT we have at least four medication names that are longer than four tokens, in the case of GGPONC 812 medications are longer than four tokens. Also in GGPONC, DEFT, and CT-EBM-SP, we can still find several entities with a span longer than ten tokens. Examples from the German data are "fettlöslichen Vitaminen" (***en:*** *"fat-soluble vitamins"*) or "orale Medikation" (***en:*** *"oral medication"*). They were both predicted correctly, however, in other cases, e.g. "schwere Beruhigungsmittel" (***en:*** *"heavy sedatives"*), this is not the case, here, simply "Beruhigungsmittel" (***en:*** *"sedative"*) would have been correct.

[24] As context, the following sentence is given: ***de:*** *"Bei der HDR-Brachytherapie werden temporär lokale Strahlenträger im Sinne einer Afterloadingtechnik in die Prostata eingebracht."*; ***en:*** *"In HDR brachytherapy, local radiation carriers are temporarily introduced into the prostate in the sense of an afterloading technique."*.

Treatment versus medication Often, there seems to be a disagreement between the terms of *treatment* (or other entity labels) and *medication*. Therapies, for example, like "chemo therapy" are dependent on the dataset, categorized in either of these categories, and therefore predicted inconsistently.

Inconsistent annotations within datasets We often encounter occurrences within datasets where the annotation might be misleading. For example, in one of the German datasets our system predicts both "Substanzen" (***en:*** *"substance"*) and "Einzelsubstanzen" (***en:*** *"single substances"*), but only the first one is a correct match.

Overlap between False Positives and False Negatives There are overall 59 expressions across all languages and datasets that are included in the FPs, but also in the FNs. Often, these mentions are from certain groups as specified above, e.g., general medication names (e.g., ***es:*** *"medicación"*), dietary supplements (e.g., ***de:*** *"Magnesium"*), or abbreviations ("ARV"). All of them, however, have a clear medical association. Their occurrence in both FPs and FNs may be a result of the different underlying guidelines or contexts, and there may be some annotation errors involved as well. However, it also demonstrates the difficulty of annotating clinical texts and creating guidelines for the annotation.

Unseen medications We take up the point of unseen mentions again. To ensure the model is not simply learning medication names by heart, we checked whether they occur in any training set. Indeed, we observe that there are several correctly predicted drugs that the model did not see during training. Examples are "Quixidar" (Quaero) and "rifampine" (DEFT). "Dexamethasone" is an interesting case: here, we can see that it was correctly predicted in both GERNERMED and GGPONC, but it never occurred like that in the training data. Instead, it was included in much longer spans, e.g., "für 3 Tage 5mg Dexamethasone" (***en:*** *"for 3 days 5mg Dexamethasone"*). Finally, examples for Spanish are "biperideno" (PharmaCoNER) or "tirofibán" (CT-EBM-SP). From this, we can conclude that context indeed plays a role when detecting medication names.

6.2.5 Summary & Conclusion

In this section, we investigated the ability of the cross- and multi-lingual transfer-learning capabilities of the XLM-R model in the context of medication detection in different languages and datasets. We fine-tuned `XLM-RoBERTa` models on monolingual, language cluster-based and multi-lingual datasets and evaluated their drug detection performance across all languages. Based on the results, we conclude that a multi-lingual model fine-tuned on all available data *does not* outperform a multi-lingual model fine-tuned on only one language when applied in the medical domain. This is not surprising and correlates with findings in the general domain.

What we find interesting, however, is that fine-tuning on clusters of similar languages *with language-weighted sampling* contributes more than fine-tuning on all available languages, except for Spanish. Further, we learned which datasets might not be helpful in this kind of experiment. For example, Ex4CDS 2.0 is probably too small to make any difference and DEFT seems to be a challenging dataset on its own. Moreover, the exact annotation of the mentions also plays an important role. We found many prediction errors to be (most likely) due to the different definitions the various corpora are based upon. This resulted in more false positives than false negatives, except for the German data, i.e., the model over-generalized on the other languages. This phenomenon needs to be investigated further since it is not yet clear how much performance is lost or gained when adding data with similar but not exact label definitions. An interesting experiment for this research direction might be monitoring the model's prediction

(un)certainty during fine-tuning and increasing or decreasing the decision boundary, similar to the approach proposed by Swayamdipta et al. (2020).

We further acknowledge the many confounding factors that still need to be investigated: In the case of English and German, for instance, it is not clear if *adding* only English or *removing* Spanish and French helped increase the performance. The Romance language cluster might be more confusing during fine-tuning for German, but the English data is very homogeneous and might not contribute much. Therefore, balanced datasets of each language with the same size could reduce this issue. On the other hand, having diverse data at hand might make the models more robust.

Also, for the next iteration of experiments, the test set should be more balanced to allow for a more fair evaluation. Currently, the analysis is dominated by the German (GGPONC v2.0) data, accounting for half of the false negatives.

In summary, apart from an entirely multi-lingual model, language-family-based models (pre-trained on more than two languages) might be an interesting alley for future fine-tuning (or pre-training) of models. They need less data than multi-lingual models, are not too specialized on only one language, and might be able to transfer knowledge more easily, particularly for specialized domains.

6.3 Detecting medical entities in the KEEPHA dataset

In this section, preliminary experiments for detecting medical entities are conducted to provide baselines for the KEEPHA dataset as described in Section 4.1. We first apply the models introduced in Section 6.2 for detecting only medication mentions. After that, we provide a simple baseline for all entities together which is also supposed to serve as annotation verification, to highlight inconsistencies in the annotations or show entities we missed.

6.3.1 Drug-only Detection

This section reports the results of the models described in Section 6.2, which were fine-tuned on medication detection only, but on datasets in four languages. Recall that we investigated multilingual models (`XLM-RoBERTa`) fine-tuned on three types of dataset combinations: One setup was fine-tuning on all available data in the languages German, English, French and Spanish (*multi-lingual*), one was fine-tuning on a single language only (*mono-lingual*) and one was fine-tuning on language clusters, i.e., German-English and French-Spanish (*clusters*).

We report the results of each dataset combination for the French and German data. Note that as before, all scores result from a majority voting of five models, each fine-tuned with a different seed. Furthermore, we used all available data in the KEEPHA datasets for testing the zero-shot models.

Results & Discussion

The results are displayed in Table 6.6. The best results for the German part of the KEEPHA dataset are achieved when using the model fine-tuned only on German, with an F_1 score of 0.77. Note that the precision of all three models is the same, but the recall changes.

For French, the best F_1 score is achieved by the model fine-tuned on all datasets in all languages, with an increase of 4 percentage points compared to the other models. These scores are better than the ones achieved on the original test sets of the model and specifically better than on the Quaero dataset ($F_1 = 64.5$ when fine-tuned on only French data and evaluated on both French test sets, $F_1 = 71.6$ when trained on all languages and evaluated only on the Quaero test set). In this context, it is interesting that the model fine-tuned on all data achieves a better score than the other combinations, implying that the other languages somehow helped detect

		lenient		
train	**test**	**precision**	**recall**	**F_1**
de	KEEPHA-de	0.85	0.71	0.77
de, en	KEEPHA-de	0.85	0.68	0.75
all	KEEPHA-de	0.85	0.69	0.76
fr	KEEPHA-fr	0.77	0.65	0.70
fr, es	KEEPHA-fr	0.79	0.63	0.70
all	KEEPHA-fr	0.81	0.69	0.74

Table 6.6: The zero-shot results (lenient) on the newly created KEEPHA corpus, but only with respect to the `drug` mentions. The first column describes the language of the data the model was previously trained on, *all* refers to German (de), English (en), French (fr), and Spanish (es). The test column describes the part of the KEEPHA data the models were tested on.

the medications. However, first of all, the French KEEPHA data is translated from German, and might still contain some rather German medication names, abbreviations, or formulations. Secondly, as observed in the previous chapter, the results on the French data were strongly influenced by the different annotation schemes with which both French corpora were annotated. This might have been balanced by including the other languages.

We do not report the results on the Japanese part of the KEEPHA dataset, since for Japanese, a different pre- and post-processing would be necessary, in particular, the way we are currently converting between `BRAT` and `BIO` needs to be modified. This is future work.

Considering that the models were fine-tuned on data from a different domain and applied without any further fine-tuning on the KEEPHA data, the results are quite good. When looking at the predictions, we again find some over-generalizations, particularly for German. For instance, "Harndrang" (***en:*** *"desire to void one's bladder"*) (which could either be a `disorder` or `function`, depending on the context) or "Speichelprobe" (***en:*** *"saliva sample"*), which would be a `test`, were both predicted as `drug`. More complicated constructions, like "Ö-Creme" (***en:*** *"estrogen creme"*), also get predicted correctly. Sometimes, `route` and `drug` entity types get confused, those depend strongly on the context. However, for French, these confusions do not seem to happen very often at first glance. Of course, the above results need to be investigated further but they already demonstrate a successful transfer across both domains and languages.

6.3.2 Baseline Models for KEEPHA

For building a first baseline model for medical Named Entity Recognition on the KEEPHA dataset, we again take `XLM-RoBERTa` (large) and run a hyper-parameter search for two experiments: First, we aim to find the best model for fine-tuning on a training set of French and German KEEPHA data, and then we would like to find the best model for the transfer between German (for fine-tuning) and French (for evaluation). The determined hyper-parameters are summarized in Appendix D.3.1. With this, we investigate the general usability of our new dataset and further test the cross-lingual transfer of annotations across languages (German to French). Furthermore, it will help us find inconsistent or missing annotations.

Using the determined hyper-parameters, the models are fine-tuned using five different seeds for initialization. The resulting models' predictions are, exactly as in Section 6.1 and Section 6.2, combined via majority vote and converted back to `BRAT` format, to be evaluated with the same evaluation script that was used for the preceding NER experiments.

Results & Discussion

The performance of the described modes is presented in Table 6.7 and Table 6.8 for the models fine-tuned on French and German and only German, respectively. The results separated by languages and the strict scores are shown in Appendix D.3.2 and Appendix D.3.3.

train set: de + fr	*lenient*		
test set: de + fr	**precision**	**recall**	**F_1**
`drug`	0.88	0.77	0.82
`disorder`	0.80	0.80	0.80
`function`	0.56	0.53	0.55
`doctor`	0.87	0.96	0.91
`other`	0.29	0.22	0.25
`change_trigger`	0.78	0.54	0.64
`anatomy`	0.65	0.85	0.73
`test`	0.59	0.63	0.61
`opinion`	0.67	0.26	0.38
`measure`	0.97	0.71	0.82
`time`	0.88	0.85	0.87
`route`	0.62	0.50	0.55
micro average	0.80	0.73	0.76
macro average	0.71	0.63	0.66

Table 6.7: Lenient results of the NER model fine-tuned on both French and German and evaluated on both French and German.

train set: de	*lenient*		
test set: de + fr	**precision**	**recall**	**F_1**
`drug`	0.87	0.85	0.86
`disorder`	0.80	0.84	0.82
`function`	0.67	0.35	0.46
`doctor`	0.90	0.96	0.93
`other`	0.31	0.28	0.29
`change_trigger`	0.75	0.35	0.47
`anatomy`	0.57	0.77	0.66
`test`	0.47	0.56	0.51
`opinion`	0.59	0.42	0.49
`measure`	0.91	0.71	0.79
`time`	0.86	0.83	0.85
`route`	0.67	0.25	0.36
micro average	0.79	0.74	0.77
macro average	0.70	0.60	0.63

Table 6.8: Lenient results of the NER model fine-tuned only on German and evaluated on both German and French.

The micro and macro average scores are comparable for both fine-tuning configurations. When fine-tuning on both German and French, the macro average F_1 score is 3 points better than when fine-tuning only on German, expressing a better result across entity types without taking the number of occurrences per type into account. However, the results per entity type differ when comparing both strategies. For example, the `drug` and `disorder` types get better results when fine-tuned on both languages, while `function` is 9 points worse.

These differences need to be investigated further by analyzing the systems' performance on each language separately. The analysis is out of the scope of this thesis, but the above experiments and their results serve as a starting point to further improve both the dataset and the models. While decent, there is still much room for improvement in detecting medical entities in user-generated texts.

6.4 Summary & Conclusion

In this chapter, different settings of medical entity extraction were explored. It started with the n2c2 shared task data, CMED, which resulted in very high scores with respect to the detection of drug mentions. However, these experiments were conducted within the same dataset, which is very uniform. Also, we could use different language- and domain-specific models to approach the task.

To investigate how the same approach works for data in different languages and also different datasets, we then presented experiments using a multi-lingual model fine-tuned on either one language, clusters based on language families, or all languages together. Models trained within language achieved the best results on the respective language's test set but were closely

followed by the cluster-based models. The multi-lingual model, in which we were mainly interested, achieved mixed results, depending on the dataset and language. Since a multi-lingual, robust model is of interest for further work, the results were analyzed in more detail. The error analysis showed that the model could highlight annotation inconsistencies and differences in annotation guidelines, that is, in the definition of what a medication mention is. However, it is not clear which of the many variables in the presented experiments are responsible for the outcomes and further investigation is necessary, to be able to improve upon the current performance. For this, a focus on one language family might help.

The same models described in the multi-lingual experiments were then used to perform a zero-shot prediction on the newly created KEEPHA data, again only for predicting `drug` mentions. We observed that for the German data, the in-language fine-tuning worked best, while for the French part of the data, fine-tuning on all datasets, that is, all languages, achieved the highest score.

Finally, in another round of experiments, we provided baseline models for medical entity detection on the KEEPHA dataset. Once fine-tuned on German and French and once fine-tuned on only German, to investigate the cross-lingual transfer capability of the multi-lingual model, this time between datasets annotated based on the same guidelines. The results are, even with a simple model like that, already promising, but further analysis is needed to see in which way the models should be improved, to achieve good results not only on `drug` or `disorder` detection, but also on entity types with fewer training examples.

Chapter 7

Conclusion & Future Work

7.1 Summary & Conclusion

This thesis investigated the cross- and multi-lingual transfer of knowledge with respect to Adverse Drug Reaction detection from user-generated texts. The heart of this work lies in creating a new multi-lingual corpus with several levels of annotations based on guidelines developed for at least three languages from different language families: German, French, and Japanese. We focused on the German and French data and described guidelines, annotation process, and the resulting dataset, which is annotated with binary labels, to distinguish between documents containing ADRs or not, entities to catch all relevant medical mentions, attributes, to define these mentions, and relations, to associate medical mentions with each other. The corpus is unique in its combination of languages and in that it defines ADRs via relations across sentences. In addition, the same guidelines were used for all data.

Since privacy is a significant concern when dealing with UGT, two approaches to gathering data without hurting peoples' privacy and their pitfalls are discussed. First, we present a prototype study to collect online users' descriptions of ADRs. This is done with the help of a questionnaire in which the participants can actively consent to sharing their data and get information about the research their data is used for. We find that users are not against sharing their data, but that an engaging survey design is important to gather responses of interest.

Second, the results of generating and translating pseudo-tweets are analyzed, highlighting potential problems this approach introduces. We further present ways in which these data are helpful and in which ways they could be improved.

The thesis then presents experiments for classifying documents containing ADRs in German. We introduce a two-step fine-tuning approach incorporating English data to counter-act the high label imbalance inherent to documents in this domain. We show that different strategies fail in the supposedly simple task of document classification and that the model with the best overall performance might not be reliable enough to identify documents describing ADRs. However, systems with lower overall scores but higher recall can still be applied to gather more relevant documents from unlabeled resources, which in turn can improve the current models.

We then conduct experiments on drug and general medical entity detection, starting with experiments on English data and our participation in the n2c2 2022 shared task. These are then extended to investigate the cross-lingual and cross-dataset performance of multi-lingual models, showing mixed results. We analyze the issues when performing entity detection across these datasets.

Using the same models as before, we then show the first zero-shot results on the newly created dataset. Finally, we provide a baseline for medical entity detection on the same dataset, which already shows promising results.

In summary, the thesis presents a further step in the direction of supporting pharmacovigilance across languages using methods from Natural Language Processing. Given the developed data and models, further data collection and cross-lingual transfer can be improved.

7.2 Outlook

During the work of this thesis, many new questions and research directions evolved. First of all, based on the insights from annotating user-generated text and the other discussed possibilities of generating data with users' consent or without private data, we would like to investigate de-identification more thoroughly, especially focusing on colloquial texts. We found that these texts, because of their inconsistent structure and creative language, are difficult to automatically de-identify. However, de-identification is of utmost importance and should be reliably applicable, and there should also be means to evaluate the completeness of the de-identification. We find both aspects to be an interesting direction of research.

Further, with respect to the published KEEPHA dataset, we plan to update the corpus frequently. First, we would like to apply the presented models for document and entity detection to new, unlabeled data from unknown distributions, e.g., different social media platforms, to find out how many relevant documents can be drawn from them. These can then be used to extend the corpus for all levels of annotations. Using the current models, the annotation time could be reduced significantly. It would moreover be interesting to apply this strategy to languages not yet included in the corpus to retrieve candidate documents, which could then be filtered further and subsequently annotated.

Gathering more positive documents would first improve the classification performance of the models and second, provide a more diverse overview of ADRs as described by laypeople. On the other side, it is also interesting to further investigate how to approach the strong label imbalance in the data. Here, methods like self-learning or adversarial training as described in Section 3.1 could be tested. Extending the pre-training of a model on unlabeled user-generated data could also be an option for improvement. From a reversed perspective, it would also be interesting to see how models trained on the KEEPHA data would do on the related datasets in Spanish, Japanese, Russian, and French.

Furthermore, as already mentioned, we would like to normalize the concepts in the KEEPHA data to one of the standard multi-lingual ontologies, e.g., UMLS. This could be an essential step to improve and facilitate communication and mutual comprehensibility between patients and physicians and to compile ADRs related to the same drug and original patient diagnosis.

Another aspect that became visible during the work on the presented data is the entanglement between mentions that describe disorders (i.e., medical signs or symptoms) and functions of the body, which, when not working properly, also often result in disorders. This is frequently the case for the descriptions of women in their menopause. We therefore would like to lay a new focus on descriptions of health issues in this phase of life, to find out more about what exactly the women are struggling with.

Related to that, we also encountered many user posts giving—at the very least—suspicious medical advice. Identifying (and maybe marking) posts that might not refer to correct medical facts is another avenue worth exploring.

Moreover, negation also plays a crucial role in detecting ADRs. We observed, for example, that many mentions of the expression "side effect" were found in the negative documents, meaning that these are either negated or happened in the past (and therefore also negated), while the patient presently feels well. This induces questions about handling these syntactic structures and, furthermore, how to model the timeline of diagnoses, medication intake, and experiences of positive and negative reactions to the medication.

Finally, another point of interest that became visible during the work with the NTCIR data is how people deal with health issues online across cultures and languages. Do they express their issues the same way or are there any differences in what is said and not said, depending on the language?

Appendix A

Additional Background Information

A.1 Language Modeling & Transformers

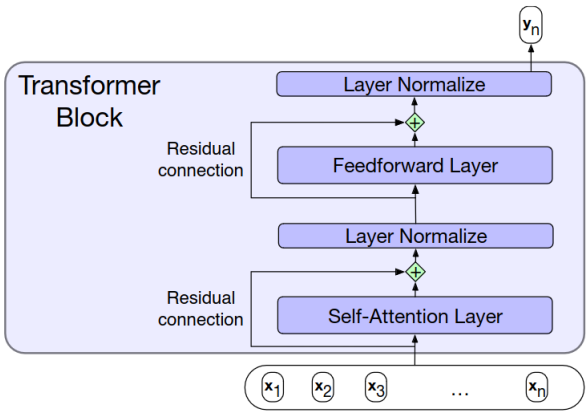


Figure A.1: A transformer block as depicted by (Jurafsky and Martin, 2023, Chapter 10, p. 216)

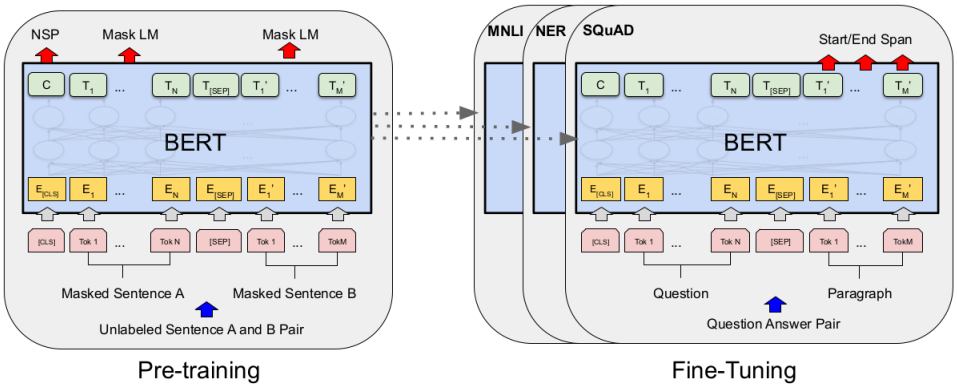


Figure A.2: The pre-training and fine-tuning procedures for BERT. Fine-tuning is done for MNLI, NER, and SQuAD. Between pre-training and fine-tuning, only the output layers differ. Image borrowed from Devlin et al. (2019).

A.1.1 Data Sizes

The phrase "a large enough corpus" for (pre-) training has been mentioned several times. But what does "large enough" mean in the context of LMs? The answer to that question depends on the task the model is trained for and the type of the model itself. For example, for training a

1000-dimensions version of the `skip-gram` model of `word2vec`, the authors used the Google News dataset[1] containing approximately 6 billion tokens, letting the model train for 2.5 days on 125 CPU cores (Mikolov et al., 2013a). Note, however, that this is not an NN.

`BERT`, on the other hand, was pre-trained on the BookCorpus (Zhu et al., 2015) and an English Wikipedia dump, containing 800 million and 2,500 million words, respectively. Pre-training of the $\texttt{BERT}_{base}$ model was done on 4 Cloud TPUs and took 4 days. Devlin et al. (2019) experimented with different numbers of Transformer blocks and self-attention heads, as well as with different hidden layer sizes. $\texttt{BERT}_{base}$ contains 12 Transformer layers, 12 heads, and a hidden layer size of 768, resulting in 110 million parameters for training. $\texttt{BERT}_{large}$ contains 24 Transformer layers, 16 heads, and a hidden layer size of 1,024 and therefore has a total of 340 million parameters. In general, for LMs and especially for (generative) LLMs, the more data are used, the better the performance (Baevski et al., 2019). The same goes for fine-tuning: The more (diverse) examples a model sees, the better it will be in most cases. According to Conneau et al. (2020), "a few hundred MiB of text data" are necessary as a minimum to train a `BERT` model.

`BERT`, and especially `BERT` in its `large` version outperformed many previously published models on the task, always using the same pre-trained model. Later, Devlin et al. (2019) also published a multi-lingual version of `BERT`, `mBERT`, trained on Wikipedia dumps in 104 languages. `mBERT` contains 12 Transformer layers, 12 heads, a hidden layer size of 768, and has 110 million parameters, i.e., the same configuration as $\texttt{BERT}_{base}$. Unfortunately, the authors do not provide the exact sizes of the Wikipedia dumps they used, but Wu and Dredze (2020) give an approximation in terms of gigabytes.

A.2 Potential Harms resulting from Language Models

With all the amazing improvements resulting from the continued improvement of Language Model one should not forget that these models can also cause harm. Pre-training on any text collection will result in models adapted to that text collection, including all stereotypes, biases and other misconceptions represented in these texts. ML models in general can replicate these factors and even reinforce them when used without care (Jurafsky and Martin, 2023).

Therefore, it is of crucial to document the architectures and training procedures of models, but also, maybe even more crucially, the datasets they are trained on. An example of how this can be done was proposed by Mitchell et al. (2019) by using *model cards*, with which each model is attached to a summary of its most important parameters. The cards should contain, for example, the training algorithms, data sources, annotation and pre-processing processes, evaluation methods, the intended use and users, and the environmental footprint. Similar approaches, i.e., giving detailed information about the data and algorithms a ML model is based on, were also proposed by Bender and Friedman (2018) and Gebru et al. (2021) for ML in general, not only NLP.

[1] Which is nowhere to be found anymore, to the best of my knowledge.

A.3 Data Sources in Biomedical NLP

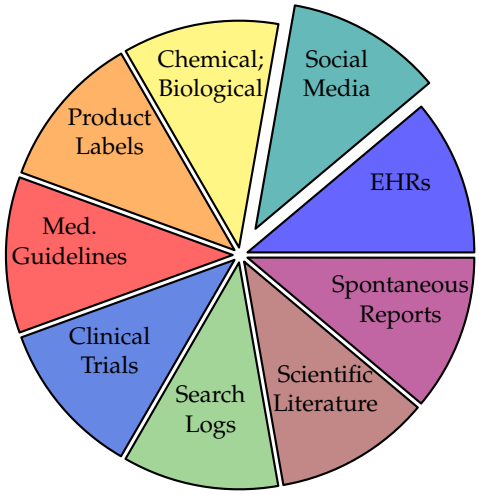


Figure A.3: The data sources that are currently used for research. The presented work focuses on data that can be summarized under "social media". Image borrowed and adapted from (Harpaz et al., 2014).

A.4 Other Resources for Biomedical NLP

There are several databases, ontologies, and concepts commonly used in different biomedical language processing which will be mentioned in the remainder of this work. For completeness, they are now briefly described.

Unified Medical Language System (UMLS) (Lindberg et al., 1993; Bodenreider, 2004) is a controlled vocabulary used in biomedical sciences, containing various terminology systems[2]. UMLS is a "Metathesaurus" mapping these different systems to each other, to provide consistency and "translations" among terms. For example, UMLS incorporates SNOMED-CT, ICD-10, and MeSH vocabularies. Often, these terms are provided in several languages as well. For many of the vocabularies, LLTs and Preferred Terms (PTs) are provided, i.e., layperson descriptions and terminology used by professionals.

SNOMED-CT was introduced as the Systematized Nomenclature of Pathology (SNOP) but later extended to the general medical field, and **C**linical **T**erms. It provides codes, terms, synonyms, and definitions as used in clinical settings for reporting[3]. It also represents the core terminology used in EHRs in various languages.

ICD-10 represents the International Statistical Classification of Diseases and Related Health Problems in its 10th revision. It is focused on, amongst other things, codes for diseases, signs and symptoms, and abnormal findings, and is managed by World Health Organization (WHO).

MeSH is used for indexing publications in the life sciences to facilitate searching. For example, it is used by PubMed to add keywords to the listed publications. Again, MeSH is organized in a hierarchy.

[2] https://www.nlm.nih.gov/research/umls/index.html
[3] https://www.nlm.nih.gov/healthit/snomedct/index.html

MedDRA (Brown et al., 1999) is a thesaurus used by the pharmaceutical industry and regulatory agencies during the process of developing, releasing, and monitoring new medications. It also contains Adverse Drug Reactions (ADRs) terminology.

CHV is a complement to UMLS, and allows the translation of complex medical terms into user-friendly language. CHV is only available in English.

SIDER (Side Effect Resource) (Kuhn et al., 2016) is a database of released medication and their ADRs, extracted from reports and medication leaflets[4] . It is only available in English.

[4] http://sideeffects.embl.de/

A.5 General Experiment Details

For all experiments, we used the Huggingface library (Wolf et al., 2020). For "traditional" machine learning models, i.e., the SVM in Chapter 5, we used the sklearn library (Pedregosa et al., 2011). Monitoring experiments was conducted with the help of Weights & Biases (Biewald, 2020) and all experiments were run on the DFKI GPU cluster. All code was written in Python 3.9.

model name	**mention**	**url**
`XLM-RoBERTa` (base)	Chapter 5, Section 6.1, Section 6.2, Section 6.3.1	`https://bit.ly/3t7A3Ga`
`BRB`	Chapter 5	`https://bit.ly/3t5bVDP`
`BioBERT`	Section 6.1	`https://bit.ly/3PT2yPF`
`PubMedBERT`	Section 6.1	`https://bit.ly/3LFKHKG`
`XLM-RoBERTa` (large)	Section 6.3.2	`https://bit.ly/3t2cVIU`

Table A.1: The models used in the different experiments, with their Hugginface URL.

Appendix B

Additional Information about the KEEPHA Corpus

B.1 Binary Annotation

B.1.1 German

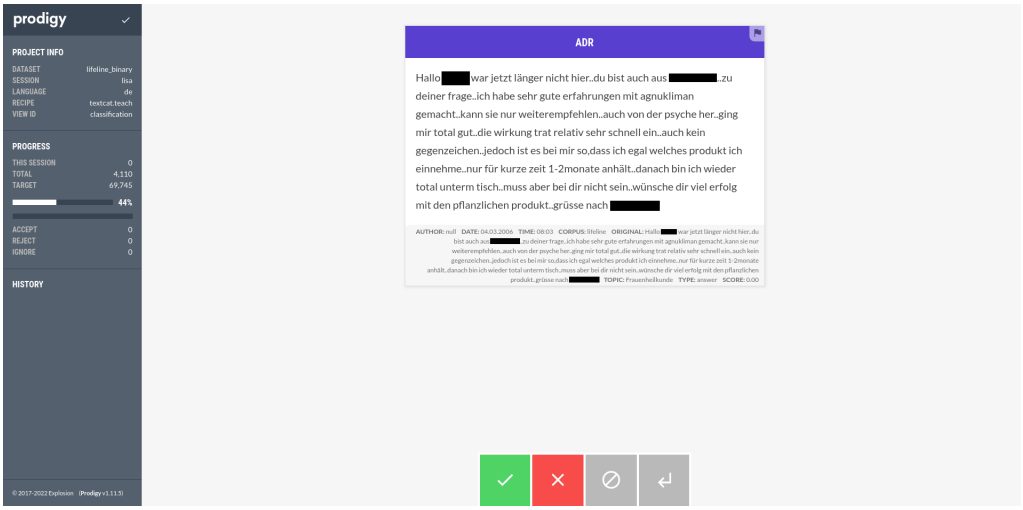


Figure B.1: The annotation interface of PRODIGY for the binary classification task. The annotators had to click the green button to mark the document as *positive* and the red button to mark it as *negative*.

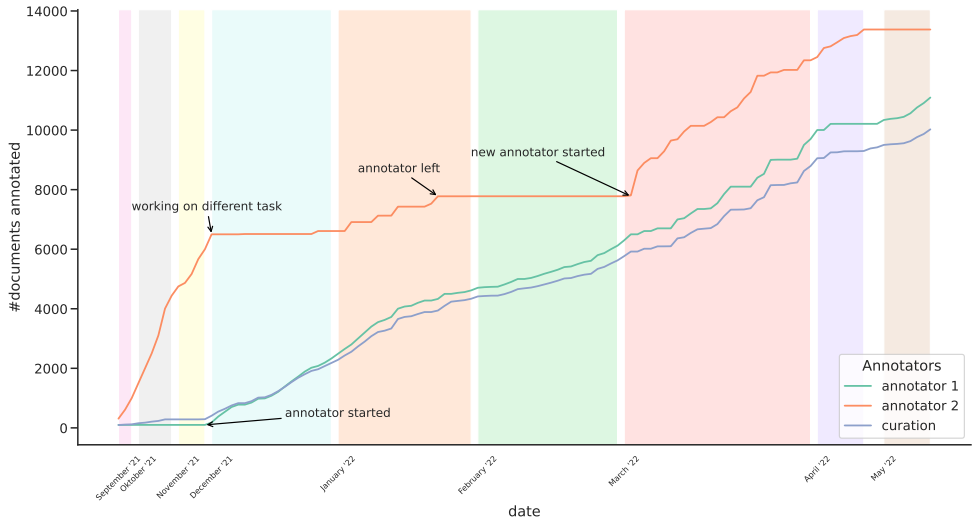


Figure B.2: The progress of annotation and consolidation of the binary annotation task over time. The colored background represents the month, while the width of the single bars shows the number of days the annotators worked on the data.

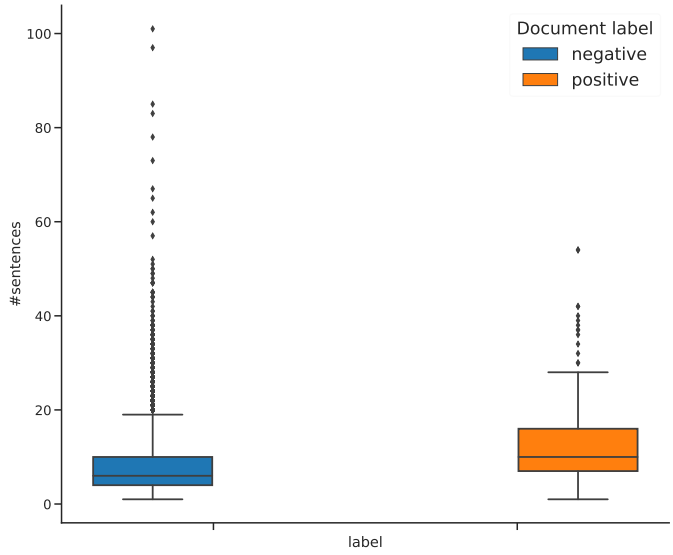


Figure B.3: The distribution of the number of sentences per document and label in the LIFELINE-DE-ALL corpus.

B.1.2 French

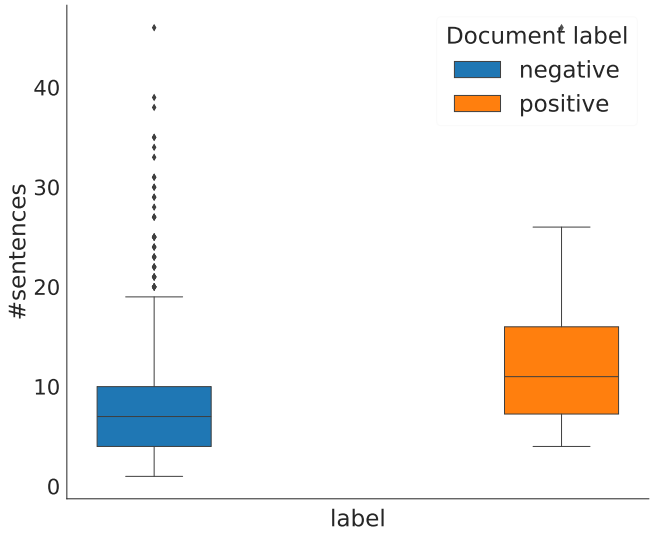


Figure B.4: The distribution of the number of sentences per document and label in the LIFELINE-1-FR.

B.2 Annotators

annotator	working language	knowledge of languages	study program	entry at DFKI	working hours
A0	de, en	German and Croatian bilingual, good knowledge of English	Pharmacy, Freie Universität Berlin, Germany	September 2021 - April 2022	10
A1	de, en	German and Russian bilingual, good knowledge of English	Pharmacy, Freie Universität Berlin, Germany	November 2021	10
A2	de, en	German and Turkish bilingual, fluent in English, basics in French and Spanish	Pharmacy, Freie Universität Berlin, Germany	March 2022	10
A3	fr, de	French (native), German (C1), English (C1)	Human Medicine, Charité Berlin, Germany	May 2023	8
A4	fr, de	Spanish (native), French (C2), German (C1), good knowledge of English	Life Science Engineering, HTW Berlin, Germany	August 2023	20

Table B.1: The background information of our annotators. Each annotator is an enrolled student and employed as student assistant at DFKI GmbH with varying working hours, depending on their availability. Each annotator earns 12,95€ per hour and they can distribute their working hours freely, with the recommendation to annotate not more than two hours continuously. The table shows the languages they were working on, the languages they know in general, their study program, the time they were hired at DFKI, and their working hours per week.

B.3 Inter-Annotator Scores

B.3.1 Entity Annotation

				relaxed		
entity type	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1**
anatomy	53	55	27	0.49	0.66	0.56
change_trigger	29	26	28	0.53	0.51	0.52
disorder	817	200	119	0.80	0.87	0.84
doctor	90	24	6	0.79	0.94	0.86
drug	553	43	39	0.93	0.93	0.93
function	125	119	57	0.51	0.69	0.59
measure	61	32	17	0.66	0.78	0.71
opinion	9	53	8	0.15	0.53	0.23
other	17	55	28	0.24	0.38	0.29
route	27	21	30	0.56	0.47	0.51
test	22	18	15	0.55	0.59	0.57
time	215	39	176	0.85	0.55	0.67
micro average	2018	685	550	0.75	0.79	0.77

Table B.2: The relaxed IAA for entity annotation in the German data.

				strict		
entity type	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1**
anatomy	45	63	34	0.42	0.57	0.48
change_trigger	21	34	36	0.38	0.37	0.38
disorder	633	384	302	0.62	0.68	0.65
doctor	89	25	7	0.78	0.93	0.85
drug	527	69	66	0.88	0.89	0.89
function	101	143	80	0.41	0.56	0.48
measure	50	43	29	0.54	0.63	0.58
opinion	4	58	13	0.06	0.24	0.10
other	17	55	28	0.24	0.38	0.29
route	25	23	32	0.52	0.44	0.48
test	19	21	18	0.48	0.51	0.49
time	167	87	227	0.66	0.42	0.52
micro average	1698	1005	872	0.63	0.66	0.64

Table B.3: The strict IAA for entity annotation in the German data.

B.3.2 Relation Annotation

						relaxed		
relation type	**head**	**tail**	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1**
caused	disorder	disorder	9	57	22	0.14	0.29	0.19
caused	disorder	function	0	5	1	0.00	0.00	0.00
caused	drug	disorder	174	105	129	0.62	0.57	0.60
caused	drug	function	0	9	8	0.00	0.00	0.00
caused	function	disorder	24	15	59	0.62	0.29	0.39
caused	function	function	1	1	3	0.50	0.25	0.33
experienced_in	disorder	anatomy	17	28	32	0.38	0.35	0.36
has_dosage	drug	measure	2	2	67	0.50	0.03	0.05
has_result	test	disorder	0	2	11	0.00	0.00	0.00
has_result	test	function	1	1	8	0.50	0.11	0.18
has_route	drug	route	2	6	44	0.25	0.04	0.07
has_time	disorder	time	6	2	63	0.75	0.09	0.16
has_time	drug	time	7	6	66	0.54	0.10	0.16
interacted_with	drug	drug	0	2	1	0.00	0.00	0.00
is_opinion_about	opinion	disorder	0	4	0	0.00	0.00	0.00
is_opinion_about	opinion	drug	2	16	16	0.11	0.11	0.11
is_opinion_about	opinion	function	0	1	0	0.00	0.00	0.00
signals_change_of	change_trigger	drug	12	16	50	0.43	0.19	0.27
treatment_for	drug	disorder	49	48	91	0.51	0.35	0.41
treatment_for	drug	function	0	14	5	0.00	0.00	0.00
micro average			306	340	676	0.47	0.31	0.38

Table B.4: The relaxed IAA for relation annotation

						strict		
relation type	**head argument**	**tail argument**	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1**
caused	disorder	disorder	5	61	22	0.08	0.19	0.11
caused	disorder	function	0	5	1	0.00	0.00	0.00
caused	drug	disorder	117	162	129	0.42	0.48	0.45
caused	drug	function	0	9	8	0.00	0.00	0.00
caused	function	disorder	12	27	59	0.31	0.17	0.22
caused	function	function	0	2	3	0.00	0.00	0.00
experienced_in	disorder	anatomy	11	34	32	0.24	0.26	0.25
has_dosage	drug	measure	1	3	67	0.25	0.01	0.03
has_result	test	disorder	0	2	11	0.00	0.00	0.00
has_result	test	function	1	1	8	0.50	0.11	0.18
has_route	drug	route	2	6	44	0.25	0.04	0.07
has_time	disorder	time	3	5	63	0.38	0.05	0.08
has_time	drug	time	6	7	66	0.46	0.08	0.14
interacted_with	drug	drug	0	2	1	0.00	0.00	0.00
is_opinion_about	opinion	disorder	0	4	0	0.00	0.00	0.00
is_opinion_about	opinion	drug	0	18	16	0.00	0.00	0.00
is_opinion_about	opinion	function	0	1	0	0.00	0.00	0.00
signals_change_of	change_trigger	drug	10	18	50	0.36	0.17	0.23
treatment_for	drug	disorder	38	59	91	0.39	0.29	0.34
treatment_for	drug	function	0	14	5	0.00	0.00	0.00
micro average			206	440	676	0.32	0.23	0.27

Table B.5: The strict IAA score for relation annotation.

B.3.3 Attribute Annotation

	strict		
attribute type	**Precision**	**Recall**	**F1**
`Negation`	0.61	0.39	0.47
`Drug_Attribute`	0.05	0.27	0.08
`Opinion_Attribute`	0.38	0.05	0.08
`Time_Attribute`	0.30	0.47	0.36
micro average	0.28	0.38	0.32

Table B.6: The strict IAA for attribute annotation in the German data.

B.4 `brat` Example

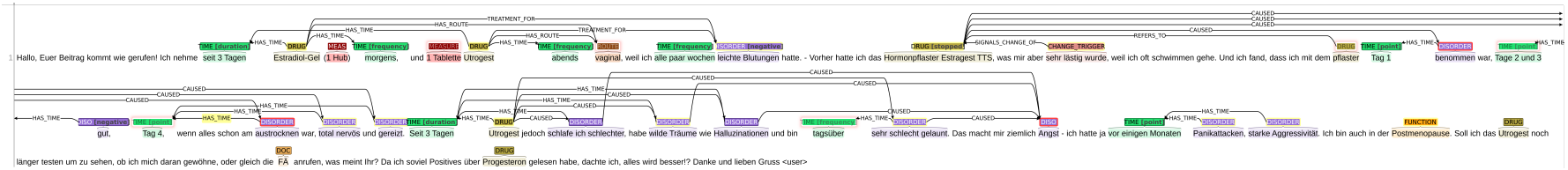
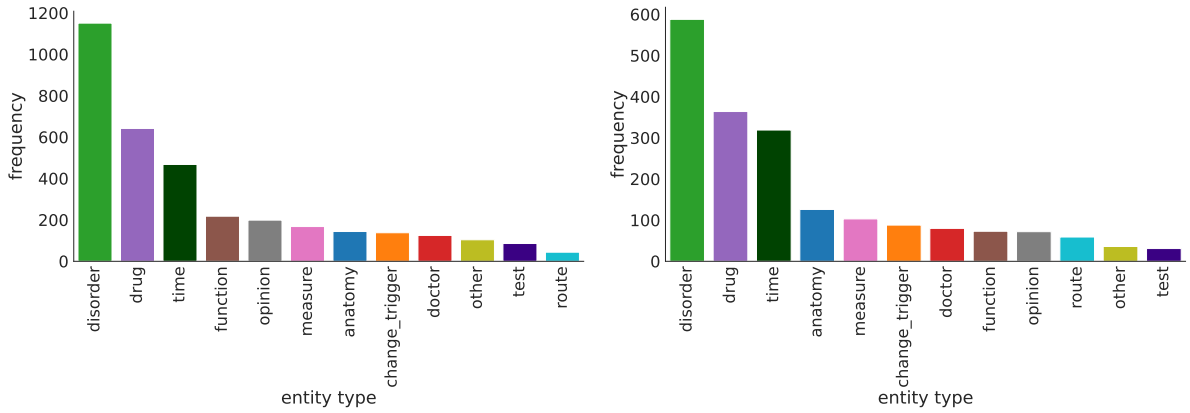


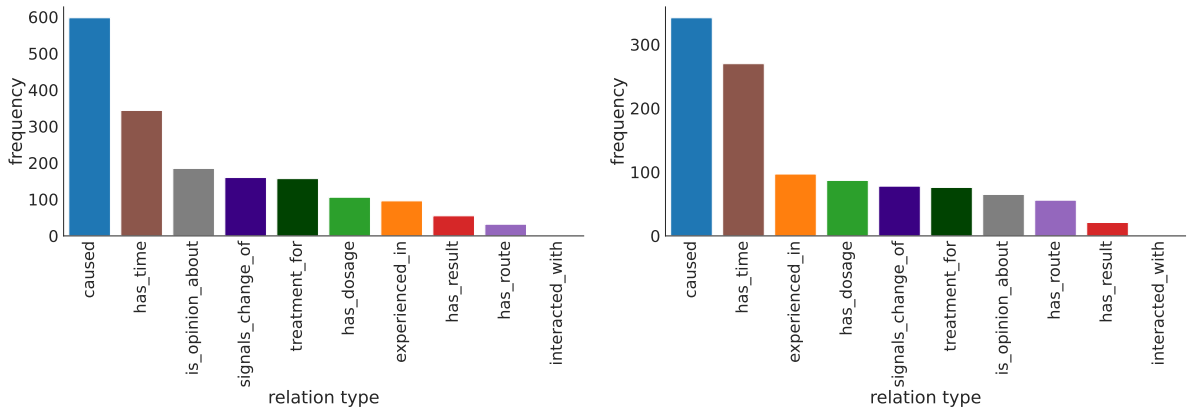
Figure B.5: An example of a `brat` annotation. This shows that annotation might get overwhelming, resulting in errors.

B.5 Dataset Statistics



(a) German data (b) French data

Figure B.6: The distribution of entity types across all documents. Note the difference in scale when comparing the two languages.



(a) German data (b) French data

Figure B.7: The distribution of relation types. Note the difference in scale when comparing the two languages.

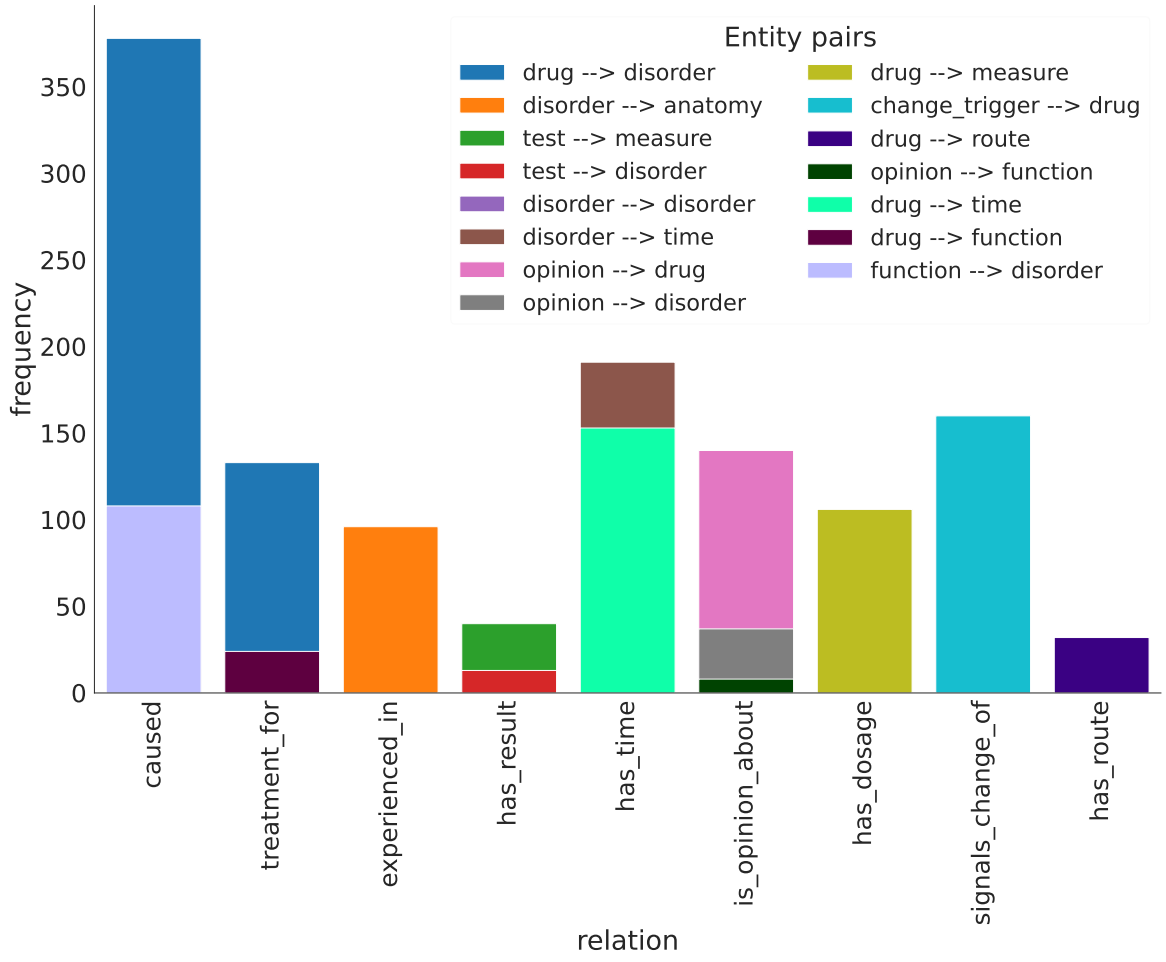


Figure B.8: The distribution of head and tail entities per relation type for the German data.

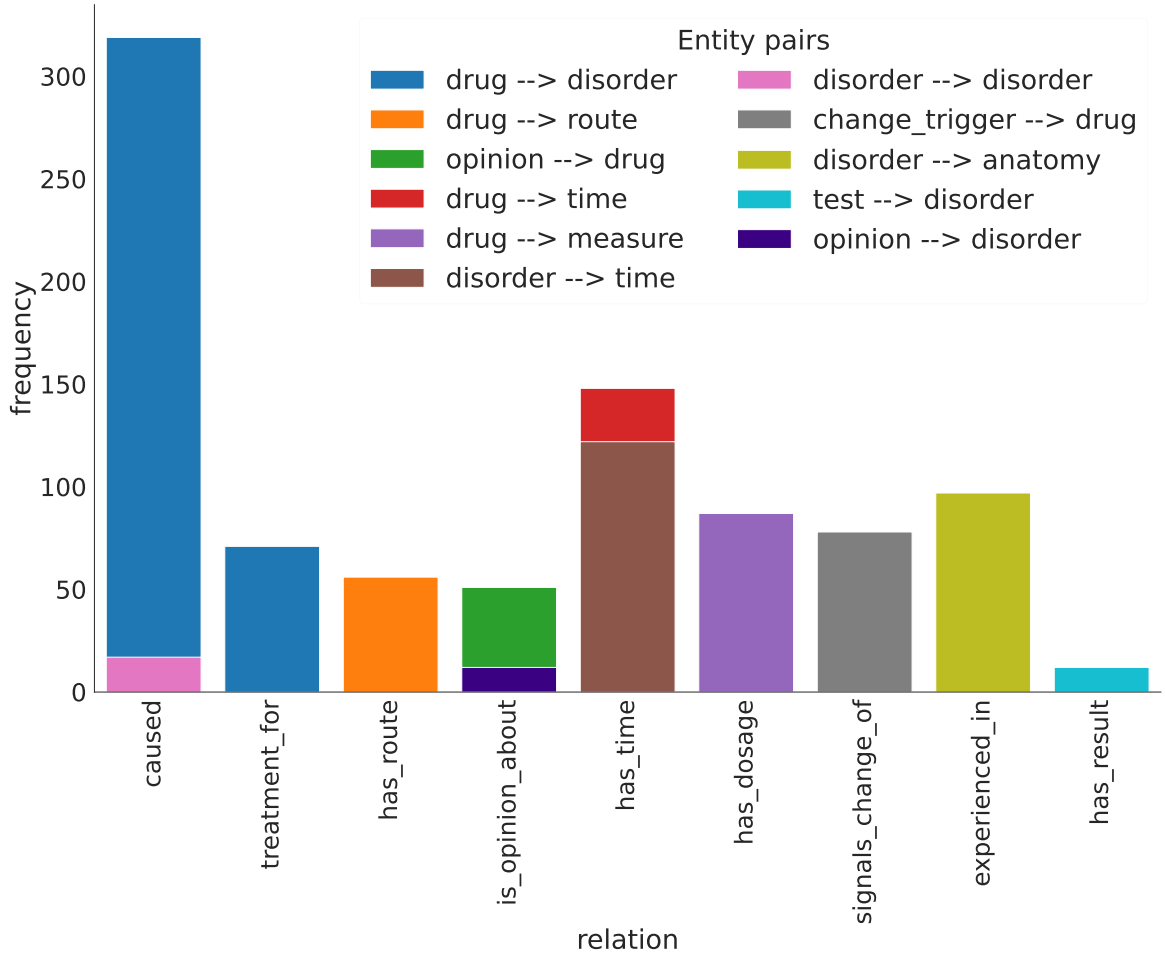
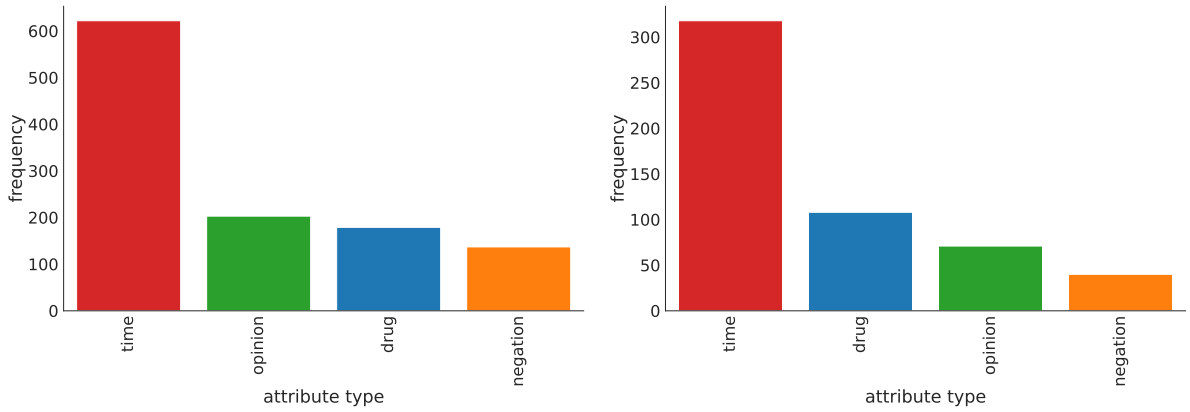


Figure B.9: The distribution of head and tail entities per relation type for the French data.



(a) German data

(b) French data

Figure B.10: The number of attribute values for each attribute type across all documents.

B.6 ADRs in German Data

drug	disorder
ad	schlecht vertragen
ad	Gewichtszunahme
ad	Libidoverlust
ads	Agressivitaet
ads	Qual
ads	eher beunruhigt
ads	Schlafentzug
ads	Gelenkschmerzen
ads	Uebelkeit
agnucaston	Zunehmen
alle bisherigen maßnahmen	Symptome verschlechtert.
antibiotikum	Hautauschlag
antibiotikum	Zittern
antibiotikum	Herzrasen
antibiotikum	hohen Blutdruck
antidepressiva	Fühle mich nicht mehr attraktiv und schön
antidepressiva	finde mich einfach wie vom anderen Stern
antidepressiva	Lust auf Sex ist einfach komplett weg
antidepressivum	Unruhe
arava	Magen-Darm Störungen
arava	starke Schmerzen
arava	blaue Flecken
arava	Entzündung
arava	Übelkeit
arimidex	Schmerzen
arimidex	Nebenwirkungen
arimidex	sehr schlimme Nebenwirkungen
asthmasprays	Nebenwirkungen
beta blocker	Müdigkeit
beta-blocker	starken Schwindel
beta-blocker	Luftnot
beta-blockers	Herzprobleme verschlimmerten sich
betablocker	Blutdruck davon so in den Keller gegangen
betablocker	Fatique Symptome
betablockern	Nebenwirkungen
bioident.	Müdigkeitsgefühl
bioident.	Benommenheit
bioident.	Dauerblutungen
bioident.	tagesmüde
bioidente hormone	enorme Magenprobleme
bioidente hormone	mir geht es eher schlechter damit, auch psychisch.
bisoprolol	Bradykard
blutdrucktabletten	Blutdruck dadurch plötzlich zu nieder wurde
calciumblocker	Schwindel
calciumblocker	Herzrasen
calciumblocker	Ruhepuls zwischen 120-160

calciumblocker	Atemnot
calciumblocker	extrem geworden
cerazette	Flecken im Gesicht
cerazette	schmerzen
cerazette	leichte Blutungen
cerazette	Stimmungsschwankungen
cerazette	3kilo runter
cerazette	esse
cerazette	essen tu ich auch net viel
cerazette	ziehen in der brust
cerazette	leichte blutungen
cerazette	vergössern sich
cerazette	Hitzewallungen
chemo	rote Wangen
chemo	unglaubliche Energie
chemo	leichten Ödemen
chemo	Nebenwirkungen
chemos	Nebenwirkungen
chlormadinon	alles durcheinander brachte
chlormadinon	Brustschmerzen
chlormadinon	Zyklus nicht richtig erholt
cipralex	fast alle Nebenwirkungen
ciprofloxacin	Darmflora geschrottet
cit	wahnsinnig gemacht
citalopram	frieren
citalopram	massive Nervosität
citalopram	Panik
citalopram	Angst
citalopram	geschwizt wie verrückt
citalopram	schreckliche Überdrehtheit
citalopram	nächtliches schwitzen
citalopram	vibrierte
citalopram	heftigen Ängsten
cliogest	vermehrten Haarwuchs
clopidogrel	sehr schlimme Nebenwirkungen
concor cor	Erektionsstörungen
concor cor	erektile Dysfunktion
cortison	Halsschmerzen
cortison	Herzrasen
cortison	sehr unwohl gefühlt
cortison	massives Verlangen nach Milch
cortison	Mir ging es nicht so gut
creme	starkes Unwohlgefühl
creme	migräneartige Kopfschmerzen
cremen	erhöhte Unruhe
cyclo progynova	nicht vertragen
doxepin	Magenbeschwerden
doxepin	Mundtrockenheit
doxepin	Panikattacke
doxepin	ganz schlimmes Brennen

doxepin	total nervos
doxepin	agressiv
escitalopram	Nebenwirkungen
escitalopram	Blutdruckkrise
escitalopram	deutlich schlechter (psychisch und körperlich)
estreva -gel	Psyche uns solche Streiche spielt
estreva -gel	alles wird nur noch schlimmer
estreva gel	vermehrte, starke Kopfschmerzen
estreva gel	Müdigkeit
estreviagel	fühlt sich wirklich schrecklich an
estreviagel	Gefühl, gleich einen Schlaganfall zu bekommen
famenita	Verdauungsprobleme
famenita	schwächer
famenita	Puddingbeine
famenita	tun weh
famenita	Schmerzen
famenita	trockenen Mund
famenita	Marathongefühl
famentinakapsel	Unruhe
famentinakapsel	viel Luft
famentinakapsel	geschlafen hab ich vorher fast besser
famentinakapsel	vom Kopf her entwas komisch
femeston conti	Blutungen
femeston conti	depressiver
fluconazol	Übelkeit
fluconazol	Erbrechen
fluconazol	nicht fähig zu arbeiten
fluconazol	nicht fähig Auto zu fahren
fluconazol	totkrank
fluconazol	Ein Gefühl wie nach Vollnarkose
fluoxetin	NW
fluoxetin	manchmal nervös
gyno	Müdigkeitsgefühl
gyno	megamüde
gyno	Benommenheit
gynokadin	Atemnot
gynokadin	niedrigen Blutdruck
gynokadin	muss mir die gesamte Haut vom Körper kratzen
gynokadin	gehen verstärkt die Haare aus
gynokadin	Haut leidet
gynokadin	unausstehlich
gynokadin	aggressiv
gynokadin	weinen
gynokadin	Schwindel
gynokadin	Dauerpinkelreiz
gynokadin	Östrogendominanz
gynokadin	nervlich dann so am Ende
gynokadin	Hitzewallungen
gynokadin	Wassereinlagerunge
gynokadingel	Beschwerden

gynokadingel	fühlt sich wirklich schrecklich an
gynokadingel	Gefühl, gleich einen Schlaganfall zu bekommen
het	überhaupt nicht mehr damit klar kam
het	Gleichgewicht unter den Hormonen fehlte
het	nicht mehr vertragen
het	Psychische Probleme
het	ganz star- ke Muskelschmerzen
het	heftigste Kopfschmerzen
het	Östrogendominanz
hormon	immer stärkeres Herzrasen
hormone	schlimme Symptome
hormone	Schwitzen
hormone	Scheidentrockenheit
hormonen	total aufgeblasen
hormonen	Spannungsgefühl
hormonen	dicke und schwere Beine
hormonpflaster estragest tts	total nervös
hormonpflaster estragest tts	benommen
hormonpflaster estragest tts	gereizt
hormonpflaster estragest tts	austrocknen
hormonpräparat	Schlimmes
hormonspirale	Myom
hormonspirale	Schmierblutungen
ibu	starkes Darmbluten
ibuprofen 600	rumpeln
immuntherapi	allergie
isoflavon-kapseln	noch mehr geschwitzt
johanneskraut	übelste Magen Darm probleme
johanneskraut	appetitosigkeit
johanniskraut	Gefühl hatte, mein allgemeiner Zustand hängt nach
johanniskraut	leicht überdreht
kalium-phosphoricum	schlimmen Nervenkrise
kalium-phosphoricum	hochtourig lief
kalium-phosphoricum	hyperaktiv lustig
kalium-phosphoricum	habe das Gefühl, es höhlt mich aus
kalium-phosphoricum	total überreizt
kalium-phosphoricum	fast luzide Träume
kalziumantagonisten	Herzrasen
kapsel von zein pharms	Kopfschmerzen
kapsel von zein pharms	Durchfall
laif 900	Weinkrämpfe
laif 900	stand irgendwie neben mir
letroblock	Hitzewellen
letroblock	ganz fürchterliche Schmerzen
lyrica	Müdigkeit
lyrica	NW
lyrica	Gewichtszunahme
lyrica	Schwitzen
lyrica	Schmerzen
magnesium	Durchfall

medikamente	voll wirre
medikamente	Nebenwirkung- en
medikamenten	Angst
medis	Bauchschmerzen
mirena	verstärkte Blutungen
mirtazapin	nahm zu sehr zu
mirtazapin	wollte nur essen
mirtazapin	dauer hungrig
mirtazepin	wie benommen
mirtazepin	steh immer noch ziemlich neben mir
mirtazepin	als sei ich nicht wirklich 'da' bzw. ich selbst
mpa gyn	periode nicht gekommen
mtx	Entzündung
mtx	Blutdruck schnellt in die Höhe
mtx	Bluthochdruck
mtx	mein Zustand hat sich auch verschlechtert
mtx	Haarausfall
mtx	Übelkeit
mtx	Nebenwirkungen sind aber geblieben
mtx	heftige Kopfschmerzen
mtx	Halsschmerzen
mtx	gerötet
mtx	Panickattacken
mtx	Angstzustände
mtx	Nebenwirkungen
mtx	Kopfschmerzen
mtx	massives Verlangen nach Milch
mönchspfeffer	heftigere Beschwerden
narkose	völlig verrückt gespielt
nat. progesteron	Dellen
nat. progesteron	bläht sich auf
nat. progesteron	Gewicht zu zunehmen
np	delliger
np	Wasser angesammelt
oekolp ovula	vermehrten Harndrang
opipramol	Benommenheit
opipramol	total neben der Spur
opipramol	Einkaufen war der pure Horror
opipramol	Benommenheitsgefühl
opipramol	müde
opipramol	brainfog
opipramol	Gefühl verrückt zu werden
opipramol	fühlte mich einfach nicht gut
opipramol	Schlaf komisch
opipramol	nicht mehr richtig wach gefühlt
opipramol	auf den Magen geschlagen
opipramol	sedierend
opipramol	Herz reagierte komisch
opipramol	mude
opipramol	Stimmungsschwankungen

opipramol	Watte im Kopf
opipramol	schnell aus der Bahn zu werfen
opipramol	seltsame Zucken
opipramol	null vertragen
opipramol	Panikattacke
opipramol	Auf und Ab
opri	stand ich völlig neben mir
opri	richtig benebelt
opri	Wie durch Watte durch den Tag
opri	Blutdruck bisschen runter
opri	noch trauriger
opripramol	innere Anspannung
opripramol	macht müde
opripramol	schwindelig
opripramol	Müdigkeit
opripramol	Stimmungseinbrüche
ovestin	vermehrten Harndrang
p	dauernd Blutungen
p-creme	Mega- blutung
pantoprazol	Blutdruck in die Höhe gegangen
pg	trockene Scheide
pille	nie vertragen
pille	dachte ich, ich drehe ab
pille	könnte ich nicht mehr schlafen
pille	Hirnarterienverschluss
pille	Problem
pille	Schlafstörungen
pille	Schwindel
pille	psychischen Nebenwirkungen
plantina	Nebenwirkungen
prog	schlimme Albträume
prog	schlimmen Träume
prog	viel Schade angerichtet
prog	chronische Müdigkeit
prog	Schwitzen
prog	Müdigkeit
prog	klatschnass geschwitzt
progesteron	Magen
progesteron	Gefühl erbrechen zu müssen
progesteron	regelrechte Übelkeit
progesteron	Druckschmerz
progesteron	Schwitzen
progesteron	Gefühl von 'aufgebläht' sein
progesteron	massive Magenprobleme
progesteron	alles wird nur noch schlimmer
progesteron	noch mehr Panik
progesteron	fing fürchterlich zu Spannen an
progesteron	aufstoßen zu müssen aber nicht zu können
progesteron	Psyche uns solche Streiche spielt
progesteron	(starke) Blutung

progesteron	Übelkeit
progesteron	extreme Brustschmerzen
progesteron	starke Kopfschmerzen
progesteroncreme	Schwindel
progesteroncreme	ganz unregelmäßigen Zyklus
progesteroncreme	Kopfschmerzen
progesteroncreme	starken Schwindel, der immer schlimmer wurde
progesteroncreme	Dauerkopfschmerzen
progesteroncreme	Schmierblutungen
progesteroncreme	Übelkeit
psorcutab beta	sehr rissig
remifemin plus	Erhöhung der Leberwerte
remifemin plus	sehr schnell schlapp
remifemin plus	ziemliche stimmungsschwankungen
reparil	dauernd übel
sanol	starken Blutungen
sanol	Schmierblutungen
schmerztabletten	hauen mir die Schuhe weg
sertralin	ewig aufgepusht
sie	Heißhungerattacken
sie	Blutdruck absackte
sie	konnte ich nicht schlafen
sie	Nebenwirkungen
sie	Ausschläge
sie	absackte Blutzucker
spirale	stärkere Blutungen
spirale	Schmierblutungen
spirale	Zwischenblutungen
tabletten	Schmerzen
tabletten	Blutungen
tamoxifen	Hitzewellen
tamoxifen	ca. 9 kg in zwei Jahren zugenommen
trigoa	Pigmentstörung
utrogest	niedrigen Blutdruck
utrogest	müde
utrogest	mischt es sich zu sehr in die Blutungen ein
utrogest	busschen matschig
utrogest	Östrogendominanz
utrogest	Wassereinlagerunge
utrogest	sehr schlecht gelaunt
utrogest	wilde Träume
utrogest	Schwindel
utrogest	Halluzinationen
utrogest	megamüde
utrogest	bleierne müde gemacht
utrogest	erwachte und fand gar keinen Schlaf mehr
utrogest	Ausschläge
utrogest	so- fortige Müdigkeit
utrogest	Müdigkeitsgefühl
utrogest	nicht vertragen

utrogest	schlafe ich schlechter
utrogest	Atemnot
utrogest	Blutdruckabfall
utrogest	nicht mehr vertrug
utrogest	Benommenheit
utrogest	nicht schlafen konnte
valium	hauen mir die Schuhe weg
venaflaxin	Unwirklichkeitsgefühle
venaflaxin	noch mehr Angst
venaflaxin	Gleichgewichtsstörungen
venaflaxin	Herzrasen
verdauungsenzyme	so schlecht
ö	Mega- blutung
östradiol	extreme Brustschmerzen
östrogen	Thrombosen
östrogen	furchtbar nervös
östrogen	geringfügige BlutdruckErhöhung
östrogen	Brustspannen
östrogen gel	Angst vor Nebenwirkungen
östrogen haltige cremes	bekomme es sofort mit der Psyche
östrogen haltige cremes	vaginal vertragen nicht
östrogene	fühlte mich damit nicht sonderlich gut
östrogene	schlafen konnte ich auch nicht vernünftig

Table B.7: The drug and disorder mentions connected by the caused relation in the German data, that is, all collected ADRs. Duplicate disorders caused by the same medication were filtered out. The table contains 368 ADRs.

B.7 ADRs in French Data

drug	disorder
ab	eu raison de moi
ad	TSH est montée en flèche
amiodarone	pique comme avec de fines aiguilles
amiodarone	je ressemble la plupart du temps à une écrevisse
analgésiques	gastro-entérite
anastrozole	une prise de poids 1,5 kg
anastrozole	enceinte de 5 mois
anastrozole	douleurs articulaires
anastrozole	syndrome du tunnel capillaire
anastrozole	maux de tête
anastrozole	ma qualité de vie en souffre
antiallergiques	somnolence
antibiotique	diarrhée
antibiotiques	problèmes de digestion
antiémétiques	somnolence
arcoxia	transpiration très forte
arimidex	douleurs articulaires
ass	brûlant
ass	rouge
ass	brûle
ass	irritée
ass	rouges
bisphosphonates	effet négatif
bisphosphonates	pas supporté
bisphosphonates	ostéonécrose
bloqueur d'acide gastrique	éruptions cutanées
bloqueur d'acide gastrique	plus marcher
bloqueur d'acide gastrique	intoxiquée
bloqueur d'acide gastrique	pas supporté
bloqueur d'acide gastrique	paralysie des muscles
bloqueur d'acide gastrique	troubles de la vue
bloqueur d'acide gastrique	brûlures
bloqueur d'acide gastrique	démangeaisons
capsule	saignements
celles-ci	saignements permanents
chimio	m'étouffe
chimio	gonflé
chimios	vagues de chaleu
chimios	prise de poids
chimios	douleurs articulaires
chlormadinone	me sentir mal
chlormadinone	augmentation de la taille
chlormadinone	mes seins ont fait exploser tous les soutiens-gorge
chlormadinone	devenue de plus en plus grosse
cimicifuga	gonflements
cimicifuga	taches rouges

cimicifuga	démangeaisons
cimicifuga	rougeurs
cimicifuga	éruptions cutanées
ciprofloxacine	terribles crises de panique
ciprofloxacine	intoxiquée
ciprofloxacine	problème
ciprofloxacine	bourdonnements
citalopram	effets secondaires
citalopram	sensation de brûlure
clopidogrel	même résultat Encore pire
comprimé	perdu du poids très rapidement
comprimé	diarrhée sévère
comprimé	vomissais
comprimé	effets secondaires importants
comprimés	grosse diarrhée
comprimés	ne supporte pas
corti	crises de transpiration
crème	je marche comme un somnifère
crème	n'arrive plus à me lever
d'aromasin	douleurs articulaires
diltiazem	me sens pas très bien
diltiazem	forte pression
diltiazem	bourdonnements devenus beaucoup plus forts
doxépine	petite sécheresse
doxépine	pression
doxépine	problème de vision de près
doxépine	prends du poids
dystologes	chuter ma tension
esomeprazol	nausées
esomeprazol	vomir
estradiol	taux d'œstrogènes descendu en dessous de 30 pg/nl
estradiol	violents vertiges
estradiol	6 pertes d'audition consécutives
estreva	me sens pas bien
estreva	fatigue de plomb
estreva	maux de tête fulgurants
famenita	complètement à côté de mes pompes
famenita	nausées
famenita	malaise total
famenita	vertiges
faminita	pertes de sang
faminita	migraines ophtalmiques
faminita	diarrhée
faminita	sensation verse un seau d'eau chaude
faminita	vertiges
faminita	maux de tête
faminita	t'abandonnaient toute molle
faminita	nausées
faminita	irrités
faminita	mal de tête fou

faminita	même problème
faminita	vomissements
fec	sensibles
fec	saignements
femoston conti	maux de tête
femoston conti	nausées
femoston conti	mauvais sommeil
fluoxétine	perdu immédiatement 3 kg
fumaderm	poussée de chaleur
gel	sensation rugueuse
gel	salive s'accumulait
gyno	pression artérielle
gyno	maux de tête
gynocadin	nausées
gynocadin	complètement à côté de mes pompes
gynocadin	vertiges
gynokadin	mal de tête fou
gynokadin	crampes d'estomac
gynokadin	perte d'appétit
gynokadin	ne supporte pas
gynokadin	vertiges
gynokadin	agitation intérieure s'aggrave
gynokadin	migraines ophtalmiques
gynokadin	me sentais déjà bizarre
gynokadin	irrités
gynokadin	pertes de sang
gynokadin	pression cardiaque
gynokadin	syndrome prémenstruel
gynokadin	nausées
gynokadin	picotements
gynokadin	surdosage
gynokadin	peur
gynokadin	maux de tête
gynokadin	t'abandonnaient toute molle
gynokadin	sensation verse un seau d'eau chaude
gynokadin	malaise total
gélule	tellement fatiguée
hormones	symptômes aggravés
insidon	un peu à côté de la plaque
iscador i	démangeaisons
iscador i	démange
kalinor	Rythme cardiaque
kliogest	saignements abondants
kyleena	long spotting
kyleena	Cycle long
kyleena	saignements
kyleena	règles beaucoup changé
kyleena	problème psychologique
l	problèmes gastro-intestinaux
l	symptômes s'aggravaient

l	nausées
laif	pleurer
laif	fortes angoisses
laif	sentais déjà pas bien
laif	pas dormi
laif 900	encore plus déprimée
lamisil	valeurs du foie légèrement plus élevées
lamisil	forte sensation de ballonnement
lanzetto	diarrhée
lanzetto	vomissements
lanzetto	nausées
lenzetto oestrogène	me sentais tellement mal
magnésium	effet
metropolol	Vertiges
metropolol	difficultés à respirer
metropolol	me sens mal
metropolol	pas d'appétit
millepertuis	très mauvaises expériences
millepertuis	me font perdre la tête
millepertuis	me tire encore plus vers le bas
millepertuis	marchais comme un zombie
millepertuis	problèmes de circulation
millepertuis laiff 900	éruptions cutanées
mirena	saignements réguliers
mirena	symptômes de la non-tolérance
mirena	saignements
mirena	muqueuses sèches
mirena	pas supportée
mirena	inflammations
mirena	nausées
mirtazapine	accumuler de l'eau capitonnées
mirtazapine	grosse somnolence
mirtazapine	troubles sont revenus
mirtazapine	prendre un peu de poids
mirtazapine	on devient blasé
molsidomin	bourdonnements devenus beaucoup plus forts
molsidomin	forte pression
molsidomin	me sens pas très bien
monuril	irritation
monuril	vulvovaginite
mtx	intolérance
mtx	transpiration
mtx	crises de transpiration
médicament	crises de panique
médicament	problèmes
médicament	hyperventilation
médicament	perdu le contrôle de ma respiration
médicaments	épuisement
médicaments	m'épuisent
médicaments	mauvais état général

métoprolol	pris plus de 10kg
nébivolol	perdu 5kg
nébivolol	légères nausées
opipramol	très secs
opipramol	pleuré
opipramol	fatiguait
opipramol	palpitations
opipramol	dans le brouillard
opipramol	respiration rapide
opipramol	dormi 11 heures de plus
opipramol	passé la moitié de la nuit debout
opipramol	trembler
opripramol	beaucoup plus de bouffées de chaleur
pansement	allergie
pansement	réaction d'hypersensibilité
pansement	cloques
pansement	brûlure
pilule normale	troubles de la vue
pilule normale	hypertension
pilule normale	saignements permanents
pilule normale	effet dévastateur
predni	effets secondaires
predni	sens à nouveau
predni	pris 7 kg
prog	nausée
prog	maux de ventre
progestérone	troubles du sommeil
progestérone	fait mal
progestérone	nausées
progestérone	ne supportes pas
progestérone	forts maux de tête
progestérone	douloureux
progestérone	douleurs mammaires s'aggravent
progestérone	taux beaucoup trop élevé
progestérone	Douleurs
progestérone	durs
progestérone	fatigue
progestérone	fatiguée
progestérone	saignements
progestérone	ne tolère pas
progestérone	syndrome prémenstruel
progestérone	tension désagréable
progestérone	tellement mal
progestérone	pique
progestérone bio-identique	me sentais très mal
progestérone bio-identique	pleurais
progésterone	dépressive
progésterone	fatiguée
progésterone	syndrome prémenstruel
prométhazine	extrêmement sèche

prométhazine	énorme somnolence
préparation	problèmes typiques
regaine	palpitations cardiaques
remifimin plus	saigner à nouveau
rimkus	fatigue
rimkus	vertiges
sepia c200	Douleurs musculaires
sepia c200	maux de tête
sepia c200	fortes douleurs abdominales
sepia c200	fortes nausées
sepia c200	douleurs
sepia c200	qui tombe
sepia c200	vertiges
sepia c200	forte transpiration avec frissons
sepia c200	forte anxiété
sulfasalazine	augmentation extrême de mes valeurs hépatiques
sulfasalazine	faire des rots
sulfasalazine	fatigue extrême
suppositoires	gonflés
suppositoires	plus d'énergie
suppositoires	gros myome
suppositoires	pris beaucoup de poids 20 kilos de trop
suppositoires	beaucoup d'appetit
tamoxifène	l'impression de rouiller
terbinafine	effets secondaires
thyrex	me sens parfois plus mal qu'avant
thyrex	pincements
thyrex	symptômes d'hyperfonctionnement
thyroxine	me sentais tellement mal
thyroxine	Pouls rapide
thyroxine	tension artérielle élevée
thyroxine	agitation
thyroxine	augmenté l'agitation
traitement hormonal substitutif	règles extrêmes
traitement hormonal substitutif	me sens pas vraiment bien non plus
trevelor	pas supporté
trevelor	encore plus fébrile
trevelor	pouls plus en plus élevé
trevilor	bouffées de chaleur
trimpramine	énorme somnolence
trimpramine	extrêmement sèche
tromcardin complex	diarrhée
utrogest	Etourdissements
utrogest	me sens pas bien
utrogest	douleurs abdominales
utrogest	ne supporte pas
utrogest	pertes de sang constantes
utrogest	douleurs musculaires extrêmes
utrogest	nausées
utrogest	très nerveuse

utrogest	contractent
utrogest	dormi que deux heures
utrogest	nausées massives
utrogest	me sens toute gonflée
utrogest	fatigue
utrogest	diarrhée terrible
valériane	très mauvaises expériences
valériane	problèmes de circulation
valériane	me font perdre la tête
valériane	me réveille
vaseline	pustules et les brûlures sont plus présentes
vaseline	supporte pas
œstrogène	douloureux
œstrogène	tellement mal
œstrogène	pique
œstrogène	durs
œstrogène	fait mal

Table B.8: The drug and disorder mentions connected by the caused relation in the French data, that is, all collected ADRs. Duplicate disorders caused by the same medication were filtered out. The table contains 313 ADRs.

B.8 User Study – Results

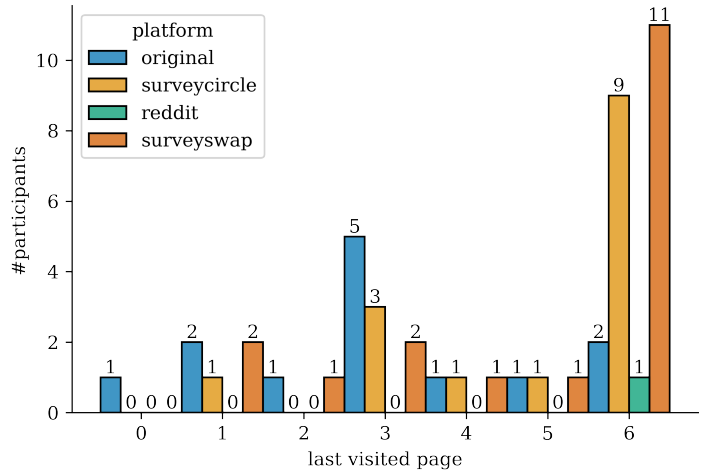


Figure B.11: The distribution of responses per platform. "original" refers to the original URL directly going to LimeSurvey.

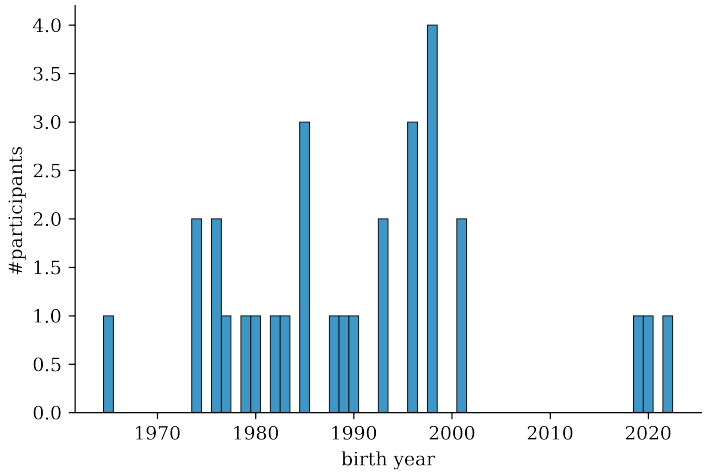


Figure B.12: The birth year distribution as provided by the participants.

B.9 NTCIR Data Validation Measures

measure	en	de	fr	total
length ratio	172	284	279	735
semantic similarity	292	313	306	911
token alignment	110	88	311	509
back-translation + token alignment	178	180	168	526
across metrics	38	64	55	157

Table B.9: The resulting numbers of found outliers per metric and language used, as well as the overall number per metric. The bottom row shows the number of samples flagged by at least three of the four measures. en: English, de: German, fr: French.

Appendix C

Additional Document Classification Results

C.1 Source Language Model Results

		negative			*positive*			*macro average*			
model	**seed**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**AUC**
XLM-R$_1$	78	66.67	71.26	68.89	92.42	90.77	91.59	79.55	81.02	80.24	81.02
XLM-R$_2$	99	68.57	55.17	61.15	88.95	93.45	91.15	78.76	74.31	76.15	74.31
XLM-R$_3$	227	62.77	67.82	65.19	91.49	89.58	90.53	77.13	78.70	77.86	78.70
XLM-R$_4$	409	66.25	60.92	63.47	90.09	91.96	91.02	78.17	76.44	77.24	76.44
XLM-R$_5$	422	70.77	52.87	60.53	88.55	94.35	91.35	79.66	73.61	75.94	73.61
XLM-R$_6$	482	64.89	70.11	67.40	92.10	90.18	91.13	78.50	80.15	79.27	80.15
XLM-R$_7$	485	59.48	79.31	67.98	94.14	86.01	89.89	76.81	82.66	78.94	82.66
XLM-R$_8$	841	61.22	68.97	64.86	91.69	88.69	90.17	76.46	78.83	77.52	78.83
XLM-R$_9$	857	67.90	63.22	65.48	90.64	92.26	91.45	79.27	77.74	78.46	77.74
XLM-R$_{10}$	910	71.43	63.22	67.07	90.75	93.45	92.08	81.09	78.34	79.58	78.34
	mean	66.00	65.29	65.20	91.08	91.07	91.03	78.54	78.18	78.12	78.18
	std	3.94	7.89	2.81	1.67	2.55	0.67	1.45	2.82	1.44	2.82

Table C.1: Source language data (English): results for XLM-RoBERTa in precision (**P**), recall (**R**) and F_1 score per class (*negative* and *positive*) and macro-averaged. The models have the same configuration and are trained and tested on the exact same data, but have a different seed for initialization. Support for the negative class: 87, support for the positive class: 336

		negative			*positive*			*macro average*			
model	**seed**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**AUC**
$\texttt{BRB}_1$	78	75.41	52.87	62.16	88.67	95.54	91.98	82.04	74.20	77.07	74.20
$\texttt{BRB}_2$	99	68.82	73.56	71.11	93.03	91.37	92.19	80.92	82.47	81.65	82.47
$\texttt{BRB}_3$	227	66.67	73.56	69.95	92.97	90.48	91.70	79.82	82.02	80.82	82.02
$\texttt{BRB}_4$	409	60.16	85.06	70.48	95.67	85.42	90.25	77.91	85.24	80.36	85.24
$\texttt{BRB}_5$	422	64.36	74.71	69.15	93.17	89.29	91.19	78.76	82.00	80.17	82.00
$\texttt{BRB}_6$	482	73.26	72.41	72.83	92.88	93.15	93.02	83.07	82.78	82.92	82.78
$\texttt{BRB}_7$	485	62.50	63.22	62.86	90.45	90.18	90.31	76.47	76.70	76.59	76.70
$\texttt{BRB}_8$	841	61.22	68.97	64.86	91.69	88.69	90.17	76.46	78.83	77.52	78.83
$\texttt{BRB}_9$	857	63.54	70.11	66.67	92.05	89.58	90.80	77.80	79.85	78.73	79.85
$\texttt{BRB}_{10}$	910	78.33	54.02	63.95	88.98	96.13	92.42	83.66	75.08	78.18	75.08
	mean	67.43	68.85	67.40	91.96	90.98	91.40	79.69	79.92	79.40	79.92
	std dev	6.32	9.79	3.79	2.11	3.23	1.01	2.64	3.64	2.10	3.64

Table C.2: Source language data (English): results for `BioRedditBERT` in precision (**P**), recall (**R**) and F_1 score per class (*negative* and *positive*) and macro-averaged. The models have the same configuration and are trained and tested on the exact same data, but have a different seed for initialization. Support for the negative class: 87, support for the positive class: 336

model	**data**	**learning rate**	**batch size**	**freeze**	**train sampler**
`XLM-R`	English	0.00001056	7	1	random
`BRB`	English	0.00001584	8	1	random
`XLM-R`	German (full)	0.00001056	7	0	weighted

Table C.3: Specifications of the best models. The first and second lines correspond to the basis for the few-shot experiments where we trained 10 versions, the bottom one is XLM-RoBERTa again fine-tuned on the full German dataset. For the first two, a random sampler and freezing all layers except the classifier worked best, while not freezing any layers and using a weighted training sampler achieved the best performance for the third model.

C.2 Results on Negative Class

			negative class			*macro average*			
model	**method**	**target data**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**AUC**
SVM	full	all	98.73	86.92	92.5	54.49	72.03	54.92	72.03
SVM	per_class	10	99.34 ± 1.01	35.52 ± 20.03	49.48 ± 23.39	51.36 ± 0.7	60.62 ± 7.64	27.99 ± 11.88	60.62 ± 7.64
SVM	per_class	40	**99.90 ± 0.22**	22.24 ± 4.73	36.17 ± 6.67	51.57 ± 0.14	60.64 ± 2.23	21.22 ± 3.48	60.64 ± 2.23
SVM	add_neg	10 + 200 neg	98.27 ± 0.17	87.25 ± 3.33	92.4 ± 1.77	53.11 ± 0.56	64.1 ± 2.64	52.78 ± 1.38	64.1 ± 2.64
SVM	add_neg	40 + 400 neg	98.98 ± 0.33	71.63 ± 2.82	83.08 ± 1.8	52.57 ± 0.3	71.53 ± 3.68	47.21 ± 0.68	71.53 ± 3.68
BRB	zero-shot	-	97.55	99.00	98.27	54.33	51.88	52.47	51.88
XLM-R	zero-shot	-	99.77	54.42	70.42	52.48	74.83	40.13	74.83
XLM-R	full	all	98.16 ± 0.19	**99.43 ± 0.23**	**98.79 ± 0.07**	**77.9 ± 3.55**	64.00 ± 3.68	**68.15 ± 3.33**	64.00 ± 3.68
XLM-R	per_class	10	99.15 ± 0.91	66.45 ± 9.91	79.21 ± 6.27	52.2 ± 1.08	70.84 ± 10.88	44.48 ± 1.9	70.84 ± 10.88
XLM-R	per_class	40	99.71 ± 0.13	61.34 ± 5.95	75.82 ± 4.69	52.87 ± 0.48	77.34 ± 3.65	43.58 ± 3.11	77.34 ± 3.65
XLM-R	add_neg	40 + 100 neg	98.22 ± 0.73	94.5 ± 8.62	96.12 ± 4.46	62.29 ± 7.67	63.44 ± 11.33	57.97 ± 6.94	63.44 ± 11.33
XLM-R	add_source	10 + 100 neg + 200 source	99.39 ± 0.27	75.99 ± 4.5	86.07 ± 2.84	53.84 ± 0.36	**78.95 ± 2.67**	50.55 ± 1.99	**78.95 ± 2.67**
XLM-R	add_source	40 + 300 neg + 300 source	98.72 ± 0.43	90.91 ± 4.82	94.59 ± 2.4	57.28 ± 3.1	72.6 ± 6.7	58.57 ± 2.8	72.6 ± 6.7
BRB	per_class	10	98.03 ± 0.12	75.54 ± 5.29	85.25 ± 3.35	51.21 ± 0.39	58.72 ± 1.98	46.58 ± 2.14	58.72 ± 1.98
BRB	per_class	40	99.14 ± 0.23	56.59 ± 8.68	71.72 ± 7.39	51.94 ± 0.35	68.77 ± 3.35	40.33 ± 4.2	68.77 ± 3.35
BRB	add_neg	40 + 100 neg	97.67 ± 0.2	99.00 ± 1.01	98.33 ± 0.4	59.79 ± 10.38	54.26 ± 3.86	54.66 ± 4.12	54.26 ± 3.86
BRB	add_source	40 + 100 neg + 200 source	97.94 ± 0.11	98.23 ± 0.48	98.08 ± 0.23	61.07 ± 2.83	59.59 ± 2.06	60.16 ± 2.08	59.59 ± 2.06

Table C.4: Target language (German): results of the best runs for every scenario and for the negative class. We excluded those with an F_1 of 0.0 for the positive class. BRB = BioRedditBERT, XLM-R = XLM-RoBERTa. **P** is precision, **R** is recall and **F_1** is F_1 score.

C.3 Experimental Setup

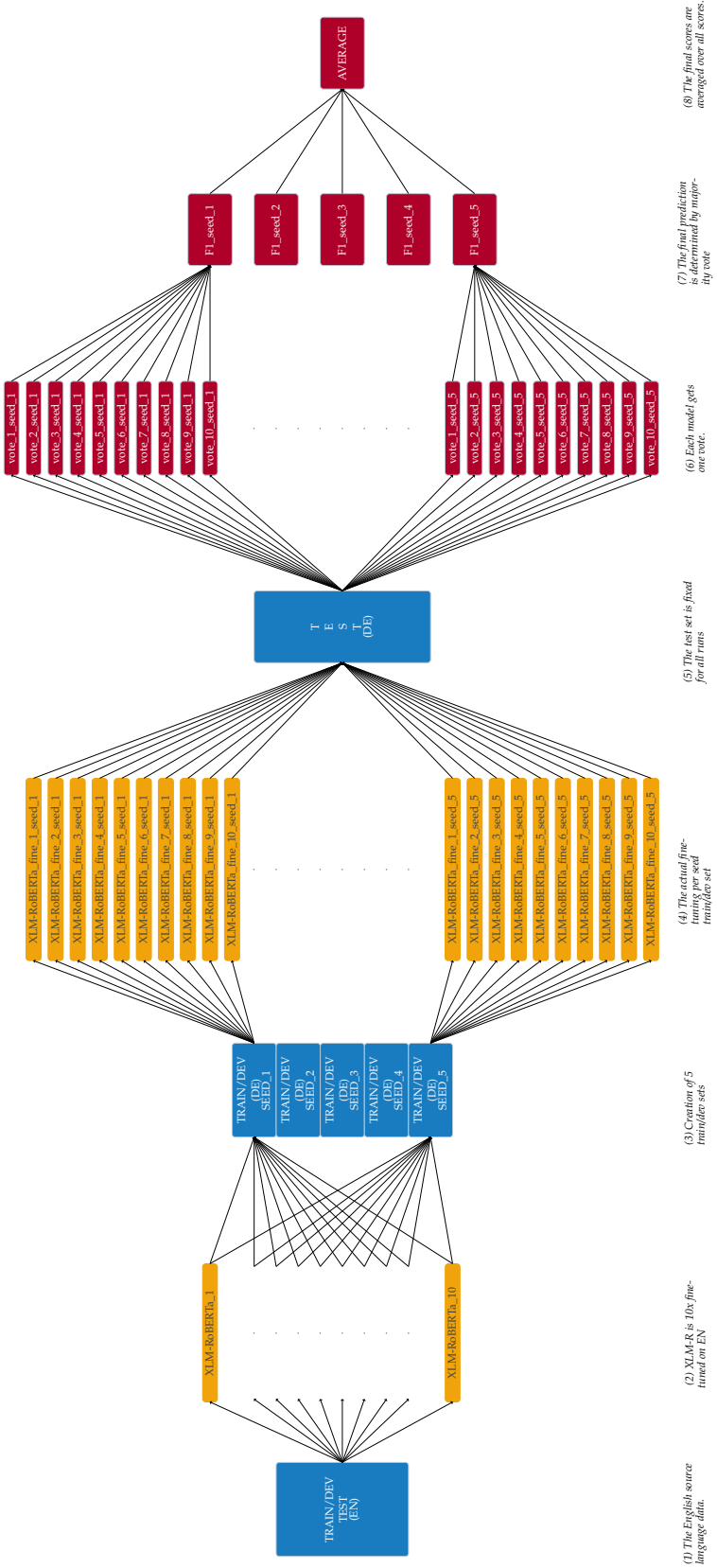


Figure C.1: The setup for the few-shot experiments: (1 + 2) We fine-tune 10 XLM-R models on the English source language data (fine-tuning 1). (3) Then, we choose 5 seeds and create 5 train/dev sets, from which we sample the shots. (4) We fine-tune (fine-tuning 2) each XLM-R model on each seed data, obtaining 10 XLM-R_fine models for every seed. (5) For every seed, each model is applied to the test set, and (6) we vote on the final results. (7) We obtain 5 results, one for every seed (F1-scores etc). (8)Those 5 results per setting are averaged.

Appendix D

Additional Information on Cross-Lingual NER

D.1 Cross-lingual Drug Detection

D.1.1 Original Labels in the Datasets

language	dataset	label
de	BRONCO150	MEDICATION
	GERNERMED	Drug
	GGPONC 2.0	Substance
	Ex4CDS 2.0	Medication
en	CMED	Dispostion/NoDisposition/Undetermined/Drug
fr	Quaero	CHEM
	DEFT	substance
es	PharmaCoNER	NORMALIZABLES/NO_NORMALIZABLES
	CT-EBM-SP	CHEM

Table D.1: The original labels per dataset.

D.1.2 Dataset Statistics

			Medication entities			
Language	**Dataset**	**# entities**	**# train**	**# dev**	**# test**	**# total**
de	BRONCO150	8,760	959	338	333	1,630
	GERNERMED	4,722	572	423	455	1,450
	GGPONC 2.0	219,711	16,473	3,832	3,366	23,671
	Ex4CDS 2.0	2,284	62	19	17	98
						26,849
en	CMED	8,993	6,196	1,033	1,764	8,993
						8,993
fr	Quaero	16,233	1,075	1,238	1,227	3,540
	DEFT	16,331	918	297	131	1,346
						4,886
es	PharmaCoNER	7,624	2,328	1,137	983	4,448
	CT-EBM-SP	46,699	5,577	1,840	1,807	9,224
						13,673
Total		331,357	34,160	10,157	10,083	54,400

Table D.2: The columns show the language, dataset name, number of annotated entity mentions overall, and the number of *medication* mentions in training, development, and test sets.

D.1.3 Model Fine-tuning Parameters

All models were trained using batches of size 8, 5,000 warm-up steps, and a weight decay of 0.002. The seeds used for the different runs were 42, 712, 9721, 26747, and 424881. Hyperparameters were determined using the Weights & Biases[1] framework. Each chunk of data contained a maximum of 26 sentences. The sentence split was done using the original BRAT scripts as described in the pre-processing section. The used learning rates for each model (ensemble) is shown below in Table D.3

model	**learning rate**
de	9.98e-6
en	9.98e-5
fr	9.98e-5
es	9.98e-5
de, en	9.98e-6
fr, es	9.98e-6
all	9.98e-6

Table D.3: The learning rates of the different models

[1] https://wandb.ai/

D.1.4 Complete Results for Cross-lingual Drug Detection

		strict			*lenient*		
train	**test**	**precision**	**recall**	**F_1**	**precision**	**recall**	**F_1**
	all	58.9	65.3	61.9	73.3	81.3	77.1
	de	65.6	67.0	66.3	85.6	87.4	**86.5**
de	en	61.6	78.0	68.8	68.7	87.0	76.8
	fr	47.9	48.1	48.0	57.3	57.5	57.4
	es	52.9	63.0	57.5	67.3	80.1	73.1
	all	61.7	52.5	56.7	74.0	63.0	68.1
	de	44.9	41.5	43.1	64.6	59.8	62.1
en	en	93.4	90.6	92.0	96.3	93.4	**94.9**
	fr	52.9	35.9	42.8	61.0	41.4	49.3
	es	70.4	52.9	60.4	78.5	59.0	67.4
	all	59.8	51.2	55.2	75.2	64.4	69.4
	de	49.0	41.2	44.7	75.6	63.5	69.1
fr	en	68.8	62.0	65.2	75.2	67.8	71.3
	fr	58.9	54.7	56.7	67.1	62.2	**64.5**
	es	70.7	57.6	63.5	79.1	64.5	71.1
	all	65.7	60.2	62.8	79.2	72.5	75.7
	de	50.8	46.1	48.3	75.7	68.8	72.1
es	en	74.0	62.9	68.0	80.4	68.4	73.9
	fr	54.4	47.6	50.8	63.2	55.4	59.1
	es	86.5	85.3	85.9	90.1	88.9	89.5

Table D.4: Results of models trained on the single languages. The evaluation scores are reported as micro scores over all test set samples and separated by language. Best scores are marked in bold font. The best score when training on one language and evaluating on all languages is underlined.

		strict			*lenient*		
train	**test**	**precision**	**recall**	**F_1**	**precision**	**recall**	**F_1**
all	all	73.1	75.0	74.0	83.8	86.0	84.8
	de	64.2	66.2	65.2	84.2	86.9	85.5
	en	87.7	92.2	89.9	90.7	95.4	93.0
	fr	58.9	52.7	55.6	66.7	59.6	63.0
	es	82.4	87.9	85.1	85.7	91.4	88.4

Table D.5: Results of the multi-lingual model trained and evaluated on all languages.

			lenient		
train	**language**	**test**	**precision**	**recall**	**F_1**
all	de	BRONCO150	84.5	88.8	86.6
		GERNERMED	94.4	88.6	91.4
		GGPONC 2.0	83.0	86.8	84.8
		Ex4CDS 2.0	71.4	29.4	41.7
	en	CMED	90.7	95.4	93.0
	fr	DEFT	18.6	56.8	28.1
		Quaero	88.9	59.9	71.6
	es	CT-EBM-SP	92.1	92.9	92.5
		PharmaCoNER	75.5	88.5	81.5

Table D.6: Results separated by dataset. The model was trained on all languages (`all`).

			strict			*lenient*		
train	**language**	**test**	**precision**	**recall**	**F_1**	**precision**	**recall**	**F_1**
all	de	BRONCO150	79.3	83.4	81.3	84.5	88.8	86.6
		GERNERMED	85.9	80.6	83.2	94.4	88.6	91.4
		GGPONC 2.0	60.2	63.0	61.6	83.0	86.8	84.8
		Ex4CDS 2.0	42.9	17.7	25.0	71.4	29.4	41.7
	en	CMED	87.7	92.2	89.9	90.7	95.4	93.0
	fr	DEFT	16.3	49.6	24.5	18.6	56.8	28.1
		QUAERO	78.6	53.0	63.3	88.9	59.9	71.6
	es	CT-EBM-SP	88.3	89.0	88.7	92.1	92.9	92.5
		PharmaCoNER	73.1	85.8	78.9	75.5	88.5	81.5

Table D.7: Result of the multi-lingual model separated by dataset, showing both strict and lenient scores.

		strict			*lenient*		
train	**test**	**precision**	**recall**	**F_1**	**precision**	**recall**	**F_1**
de, en	all	68.5	66.3	67.4	81.7	79.1	80.4
fr, es	all	61.5	63.9	62.7	74.3	77.3	75.8
de, en	de	66.3	66.6	66.5	86.3	86.6	86.4
de, en	en	89.5	91.6	90.5	92.6	94.8	93.7
de, en	fr	51.8	44.9	48.1	60.3	52.2	56.0
de, en	es	65.0	60.4	62.6	76.6	71.2	73.8
fr, es	de	49.8	48.8	49.3	74.2	72.8	73.5
fr, es	en	60.3	69.7	64.7	66.7	77.1	71.5
fr, es	fr	57.8	55.3	56.5	65.3	62.6	63.9
fr, es	es	79.4	87.0	83.0	83.3	91.3	87.1

Table D.8: Results using language clusters for fine-tuning, showing both strict and lenient results.

D.1.5 Error Groups

group	de	en	fr	es
proteins	Cyclin E	Creatine Kinase	PHOSPHOMONOESTÉRASE	proteína C
chemical compounds	Dinitrotoluol	phosphate	D-glycosylamines	fósforo
abbreviations	HLA	ASA	STH	PTH
generic drug names	Medikation	pain medication		narcóticos
medical terms/tools	Gewebsflüssigkeit	Tegaderm	solution	concentrado
dietary supplements	Vitamin C	B12		calcio

Table D.9: The most noticeable error groups in the false positives with examples from each language.

group	de	en	fr	es
therapies	Sorafenibtherapie	lipid-lowering therapy	traitement antidotique	
generic drug names	Herzentlastungsmedikamente	BP meds	ANTICOAGULANTS	antitrombótica
brand names	Sab Simplex		IONSYS	McGhan
medication + route	Irinotecan (60 mg/m²)		Comprimé	
ambiguous/short mentions	B6	Mg	CE	P

Table D.10: The most noticeable error groups with examples from each language for in the analysis of false negatives.

D.2 Strict results on the KEEPHA corpus for entity type drug

only drug		*strict*		
train	**test**	**precision**	**recall**	**F_1**
de	KEEPHA-de	0.78	0.66	0.72
de, en	KEEPHA-de	0.79	0.63	0.70
all	KEEPHA-de	0.78	0.64	0.70
fr	KEEPHA-fr	0.69	0.59	0.63
fr, es	KEEPHA-fr	0.71	0.56	0.63
all	KEEPHA-fr	0.72	0.61	0.66
all	KEEPHA-ja	0.43	0.02	0.04

Table D.11: The zero-shot results (strict) on the newly created KEEPHA corpus, but only with respect to the drug mention. The first columns describes the language of the data the model was previously trained on, *all* refers to German (de), English (en), French (fr) and Spanish (es). The test column describes the part of the KEEPHA data the models were tested on. Note that the low scores for Japanese are mostly due to tokenization mismatches.

D.3 Medical Named Entity Recognition on the KEEPHA data

D.3.1 Model Fine-tuning Parameters

For NER on the KEEPHA data, we used `XLM-RoBERTa` (large). All models were trained using batches of size 16, 5,000 warm-up steps, and a weight decay of 0.002 on one NVIDIA RTX A6000. The seeds used for the different runs were 42, 712, 9721, 26747, and 424881. Hyperparameters were determined using the Weights & Biases[2] framework. The sentence split was done using the original `BRAT` scripts as described in the pre-processing section. The used learning rates for each model (ensemble) is shown below in Table D.3

model	**learning rate**	**#sentences**
de + fr	9.98e-5	5
de	9.98e-4	3

Table D.12: The learning rates and number of sentences per chunk for the two experiment settings.

D.3.2 Result of the multi-lingually fine-tuned model

train set: de + fr	*lenient*		
test set: fr	**precision**	**recall**	**F_1**
`drug`	0.88	0.69	0.77
`disorder`	0.85	0.90	0.88
`function`	1.00	0.75	0.86
`doctor`	0.78	1.00	0.88
`other`	0.00	0.00	0.00
`change_trigger`	0.88	0.54	0.67
`anatomy`	0.88	1.00	0.93
`test`	0.33	1.00	0.50
`opinion`	0.75	0.25	0.38
`measure`	1.00	0.67	0.80
`time`	0.91	0.84	0.88
`route`	0.63	0.71	0.67
micro average	0.86	0.77	0.81
macro average	0.74	0.70	0.68

Table D.13: The lenient results of the NER model trained on both French and German. The table shows the results on only the French part of the test set.

[2] `https://wandb.ai/`

train set: de + fr	*lenient*		
test set: de	**precision**	**recall**	**F_1**
drug	0.88	0.83	0.85
disorder	0.77	0.75	0.76
function	0.46	0.46	0.46
doctor	0.90	0.95	0.93
other	0.31	0.24	0.27
change_trigger	0.70	0.54	0.61
anatomy	0.44	0.67	0.53
test	0.64	0.60	0.62
opinion	0.64	0.27	0.38
measure	0.93	0.76	0.84
time	0.86	0.86	0.86
route	0.60	0.33	0.43
micro average	0.76	0.70	0.73
macro average	0.68	0.60	0.63

Table D.14: The lenient results of the NER model trained on both French and German. The table shows the results on only the German part of the test set.

train set: de + fr	*strict*		
test set: de + fr	**precision**	**recall**	**F_1**
drug	0.85	0.74	0.79
disorder	0.57	0.58	0.57
function	0.50	0.47	0.48
doctor	0.87	0.96	0.91
other	0.21	0.17	0.19
change_trigger	0.72	0.50	0.59
anatomy	0.59	0.77	0.67
test	0.41	0.44	0.42
opinion	0.20	0.08	0.11
measure	0.73	0.54	0.62
time	0.72	0.69	0.70
route	0.46	0.38	0.41
micro average	0.65	0.60	0.62
macro average	0.57	0.53	0.54

Table D.15: The strict results of the NER model trained on both French and German and tested on both French and German. The table shows the resulting scores independent of language.

train set: de + fr	*strict*		
test set: fr	**precision**	**recall**	$\mathbf{F_1}$
drug	0.83	0.66	0.73
disorder	0.66	0.70	0.68
function	1.00	0.75	0.86
doctor	0.78	1.00	0.88
other	0.00	0.00	0.00
change_trigger	0.75	0.46	0.57
anatomy	0.81	0.93	0.87
test	0.33	1.00	0.50
opinion	0.00	0.00	0.00
measure	0.88	0.58	0.70
time	0.78	0.72	0.75
route	0.63	0.71	0.67
micro average	0.74	0.66	0.70
macro average	0.62	0.63	0.60

Table D.16: The strict results of the NER model trained on both French and German and tested on French. The table shows the results of only the French part of the test set.

train set: de + fr	*strict*		
test set: de	**precision**	**recall**	$\mathbf{F_1}$
drug	0.86	0.82	0.84
disorder	0.52	0.50	0.51
function	0.38	0.38	0.38
doctor	0.90	0.95	0.93
other	0.23	0.18	0.20
change_trigger	0.70	0.54	0.61
anatomy	0.39	0.58	0.47
test	0.43	0.40	0.41
opinion	0.27	0.12	0.16
measure	0.57	0.47	0.52
time	0.67	0.67	0.67
route	0.20	0.11	0.14
micro average	0.60	0.56	0.58
macro average	0.51	0.48	0.49

Table D.17: The strict results of the NER model trained on both French and German and tested on German. The table shows the results of only the German part of the test set.

D.3.3 Result of the mono-lingually fine-tuned model

train set: de	*strict*		
test set: de + fr	**precision**	**recall**	$\mathbf{F_1}$
`drug`	0.82	0.80	0.81
`disorder`	0.61	0.64	0.63
`function`	0.67	0.35	0.46
`doctor`	0.90	0.96	0.93
`other`	0.25	0.22	0.24
`change_trigger`	0.75	0.35	0.47
`anatomy`	0.49	0.65	0.56
`test`	0.32	0.38	0.34
`opinion`	0.30	0.21	0.25
`measure`	0.53	0.41	0.47
`time`	0.67	0.64	0.66
`route`	0.17	0.06	0.09
micro average	0.64	0.60	0.62
macro average	0.54	0.47	0.49

Table D.18: Strict results of the NER model fine-tuned only on German and averaged over both languages.

train set: de	*strict*		
test set: fr	**precision**	**recall**	$\mathbf{F_1}$
drug	0.74	0.75	0.75
disorder	0.63	0.64	0.63
function	0.67	0.50	0.57
doctor	0.88	1.00	0.93
other	0.00	0.00	0.00
change_trigger	0.80	0.31	0.44
anatomy	0.67	0.71	0.69
test	0.00	0.00	0.00
opinion	0.17	0.17	0.17
measure	0.65	0.46	0.54
time	0.70	0.62	0.66
route	0.00	0.00	0.00
micro average	0.64	0.60	0.62
macro average	0.49	0.43	0.45

Table D.19: Strict results of the NER model fine-tuned only on German and tested on only the French part of the corpus.

train set: de	*strict*		
test set: de	**precision**	**recall**	$\mathbf{F_1}$
drug	0.89	0.84	0.86
disorder	0.60	0.64	0.62
function	0.67	0.31	0.42
doctor	0.90	0.95	0.93
other	0.27	0.24	0.25
change_trigger	0.71	0.38	0.50
anatomy	0.35	0.58	0.44
test	0.40	0.40	0.40
opinion	0.40	0.23	0.29
measure	0.40	0.35	0.38
time	0.64	0.67	0.66
route	0.33	0.11	0.17
micro average	0.63	0.60	0.62
macro average	0.55	0.48	0.49

Table D.20: Strict results of the NER model fine-tuned only on German and tested on only the German part of the corpus.

train set: de	*lenient*		
test set: fr	**precision**	**recall**	$\mathbf{F_1}$
drug	0.84	0.85	0.85
disorder	0.86	0.89	0.87
function	0.67	0.50	0.57
doctor	0.88	1.00	0.93
other	0.00	0.00	0.00
change_trigger	0.80	0.31	0.44
anatomy	0.80	0.86	0.83
test	0.25	1.00	0.40
opinion	0.50	0.50	0.50
measure	0.94	0.67	0.78
time	0.91	0.80	0.85
route	0.33	0.14	0.20
micro average	0.83	0.77	0.80
macro average	0.65	0.63	0.60

Table D.21: Lenient results of the NER model fine-tuned only on German and tested on only the French part of the corpus.

train set: de	*lenient*		
test set: de	**precision**	**recall**	$\mathbf{F_1}$
drug	0.90	0.86	0.88
disorder	0.77	0.82	0.80
function	0.67	0.31	0.42
doctor	0.90	0.95	0.93
other	0.33	0.29	0.31
change_trigger	0.71	0.38	0.50
anatomy	0.40	0.67	0.50
test	0.53	0.53	0.53
opinion	0.67	0.38	0.49
measure	0.87	0.76	0.81
time	0.83	0.86	0.84
route	1.00	0.33	0.50
micro average	0.77	0.73	0.75
macro average	0.72	0.60	0.63

Table D.22: Lenient results of the NER model fine-tuned only on German and tested on only the German part of the corpus.

Bibliography

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131. Cited on pages 30, 32, and 98.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022. Cited on pages 40 and 42.

Rakesh Agrawal. 1994. Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference*. Cited on page 35.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yonas Kbrom, Yunyao Li, and Huaiyu Zhu. 2016. Multilingual Information Extraction with PolyglotIE. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 268–272, Osaka, Japan. The COLING 2016 Organizing Committee. Cited on page 30.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics. Cited on page 29.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. Curated Embeddings for Low-Resourced Languages: The Case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association. Cited on page 32.

Ilseyar Alimova, Elena Tutubalina, Julia Alferova, and Guzel Gafiyatullina. 2017. A machine learning approach to classification of drug reviews in Russian. In *2017 Ivannikov ISPRAS Open Conference (ISPRAS)*, pages 64–69. Cited on pages 33, 34, 36, 43, 44, 47, and 48.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics. Cited on pages 36 and 37.

Nestor Alvaro, Yusuke Miyao, and Nigel Collier. 2017. TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. *JMIR Public Health and Surveillance*, 3(2). Cited on pages 34 and 39.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively Multilingual Word Embeddings. Cited on page 31.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kayo Waki, and Kazuhiko Ohe. 2010. Extraction of adverse drug effects from clinical records. *Studies in Health Technology and Informatics*, 160(Pt 1):739–743. Cited on pages 34, 35, and 38.

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the NTCIR-11 MedNLP-2 Task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences*, Tokyo. Cited on page 86.

Vinayak Arannil, Tomal Deb, and Atanu Roy. 2023. ADEQA: A Question Answer based approach for joint ADE-Suspect Extraction using Sequence-To-Sequence Transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 206–214, Toronto, Canada. Association for Computational Linguistics. Cited on pages 34 and 38.

Yuki Arase, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Annotation of adverse drug reactions in patients' Weblogs. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6769–6776. Cited on pages 2, 43, 44, 47, and 48.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics. Cited on pages 30 and 32.

A. R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proceedings of the AMIA Symposium*, pages 17–21. Cited on page 35.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236. Cited on page 35.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610. Cited on page 32.

Jimmy Lei Ba, Kiros Jamie Ryan, and Geoffrey E. Hinton. 2016. Layer Normalization. In *Advances in NeurIPS 2016 Deep Learning Symposium*. Cited on page 25.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven Pretraining of Self-attention Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China. Association for Computational Linguistics. Cited on page 128.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Cited on page 24.

Qiyu Bai, Qi Dan, Zhe Mu, and Maokun Yang. 2019. A Systematic Review of Emoji: Current Research and Future Perspectives. *Frontiers in Psychology*, 10:2221. Cited on page 91.

M. Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-Resource Cross-Lingual Named Entity Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7415–7423. Cited on page 30.

Marco Basaldella and Nigel Collier. 2019. BioReddit: Word Embeddings for User-Generated Biomedical NLP. *Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis*, 10:34–38. Cited on page 98.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics. Cited on pages 36 and 49.

Oliver J. Bear Don't Walk, Harry Reyes Nieva, Sandra Soo-Jin Lee, and Noémie Elhadad. 2022. A scoping review of ethics considerations in clinical natural language processing. *JAMIA open*, 5(2):ooac039. Cited on page 51.

Patrick E. Beeler, Thomas Stammschulte, and Holger Dressel. 2023. Hospitalisations Related to Adverse Drug Reactions in Switzerland in 2012–2019: Characteristics, In-Hospital Mortality, and Spontaneous Reporting Rate. *Drug Safety*. Cited on pages 1 and 3.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics. Cited on page 37.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604. Cited on pages 92 and 128.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3(1532-4435):1137–1155. Cited on page 23.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically Aware Neural Model for Named Entity Recognition in Low Resource Transfer Settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics. Cited on pages 30 and 32.

Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. Cited on pages 106 and 131.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York. Cited on page 20.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022. Cited on page 22.

Tamara Bobic, Roman Klinger, Philippe Thomas, and Martin Hofmann-Apitius. 2012. Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 35–43. Cited on page 34.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270. Cited on pages 110 and 129.

Andreea Bodnari, Aurélie Névéol, Ozlem Uzuner, Pierre Zweigenbaum, and Peter Szolovits. 2013. Multilingual named-entity recognition from parallel corpora. In *CEUR Workshop Proceedings*, volume 1179. Cited on page 41.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Cited on page 98.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P Schapranow. 2022. GGPONC 2.0 - The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3650–3660. Cited on page 110.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA. Association for Computing Machinery. Cited on pages 20 and 98.

Elliot G. Brown, Louise Wood, and Sue Wood. 1999. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety*, 20(2):109–117. Cited on page 130.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, and Tom Henighan. 2020. Language Models are Few-Shot Learners. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Cited on pages 53 and 85.

Aurel Cami, Alana Arnold, Shannon Manzi, and Ben Reis. 2011. Predicting adverse drug events using pharmacological network models. *Science Translational Medicine*, 3(114):114ra127. Cited on page 35.

Leonardo Campillos-Llanos, Adrián Capllonch-Carrión, Ana Valverde-Mateos, and Antonio Moreno-Sandoval. 2022. Clinical Trials for Evidence-Based Medicine in Spanish (CT-EBM-SP) Corpus and word-embeddings. Cited on page 111.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, ICML'93, pages 41–48, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Cited on page 33.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: The

role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):540–543. Cited on page 40.

B. Chee, K.G. Karahalios, and B. Schatz. 2009a. Social Visualization of Health Messages. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10. Cited on page 35.

Brant Chee, Richard Berlin, and Bruce Schatz. 2009b. Measuring population health using personal health messages. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009:92–96. Cited on page 35.

Brant W. Chee, Richard Berlin, and Bruce Schatz. 2011. Predicting adverse drug events from personal health messages. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2011:217–226. Cited on pages 34 and 35.

Yuxuan Chen, David Harbecke, and Leonhard Hennig. 2022a. Multilingual Relation Classification via Efficient and Effective Prompting. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1059–1075, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. Cited on pages 32 and 33.

Yuxuan Chen, Jonas Mikkelsen, Arne Binder, Christoph Alt, and Leonhard Hennig. 2022b. A Comparative Study of Pre-trained Encoders for Low-Resource Named Entity Recognition. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 46–59, Dublin, Ireland. Association for Computational Linguistics. Cited on page 33.

Nancy Chinchor. 1992. MUC-4 Evaluation Metrics. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*. Cited on page 13.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 Evaluation Metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*. Cited on page 14.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving Multilingual Models with Language-Clustered Vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online. Association for Computational Linguistics. Cited on page 32.

D. V. Cicchetti and A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558. Cited on page 15.

CLEF. 2016. A brief history of cross language.... Cited on page 29.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46. Cited on page 15.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. 2008. BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941. Cited on page 33.

Carlo Combi, Margherita Zorzi, Gabriele Pozzani, Ugo Moretti, and Elena Arzenton. 2018. From narrative descriptions to MedDRA: Automagically encoding adverse drug reactions. *Journal of Biomedical Informatics*, 84:184–199. Cited on pages 34 and 36.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. Cited on pages 29, 31, 98, and 128.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. Cited on pages 30 and 31.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics. Cited on pages 30 and 31.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLOS Computational Biology*, 9(2):e1002854. Cited on page 43.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407. Cited on page 22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA. Association for Computational Linguistics. Cited on pages 26, 27, 29, 30, 31, 99, 127, and 128.

Pengjie Ding, Lei Wang, Yaobo Liang, Wei Lu, Linfeng Li, Chun Wang, Buzhou Tang, and Jun Yan. 2020. Cross-Lingual Transfer Learning for Medical Named Entity Recognition. In *Database Systems for Advanced Applications*, Lecture Notes in Computer Science, pages 403–418, Cham. Springer International Publishing. Cited on page 41.

Xin Dong and Gerard de Melo. 2019. A Robust Self-Learning Framework for Cross-Lingual Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6306–6310, Hong Kong, China. Association for Computational Linguistics. Cited on pages 30 and 32.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard de Melo. 2020. Leveraging Adversarial Training in Self-Learning for Cross-Lingual Text Classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1541–1544, New York, NY, USA. Association for Computing Machinery. Cited on page 32.

Andres Duque, Juan Martinez-Romo, and Lourdes Araujo. 2016. Can multilinguality improve Biomedical Word Sense Disambiguation? *Journal of Biomedical Informatics*, 64:320–332. Cited on page 41.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. *Proceedings of NAACL-HLT*. Cited on page 42.

Abteen Ebrahimi and Katharina Kann. 2021. How to Adapt Your Pretrained Multilingual Model to 1600 Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics. Cited on pages 30 and 32.

I. Ralph Edwards and Jeffrey K. Aronson. 2000. Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet*, 356(9237):1255–1259. Cited on page 1.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient Multi-lingual Language Model Fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China. Association for Computational Linguistics. Cited on page 32.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211. Cited on page 21.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250. Cited on pages 30 and 33.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587. Cited on page 22.

Yimin Fan, Yaobo Liang, Alexandre Muzio, Hany Hassan, Houqiang Li, Ming Zhou, and Nan Duan. 2021. Discovering Representation Sprachbund For Multilingual Pre-Training. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 881–894, Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 32.

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications.* MIT Press. Cited on page 22.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual Open Relation Extraction Using Cross-lingual Projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics. Cited on page 32.

John Rupert Firth. 1957. A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford. Cited on page 21.

J. L. Fleiss. 1975. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31(3):651–659. Cited on page 15.

Elizabeth Ford, Scarlett Shepherd, Kerina Jones, and Lamiece Hassan. 2021. Toward an Ethical Framework for the Text Mining of Social Media for Health Research: A Systematic Review. *Frontiers in Digital Health*, 2:592237. Cited on pages 48, 49, 50, and 51.

Johann Frei and Frank Kramer. 2022. GERNERMED: An open German medical NER model. *Software Impacts*, 11:100212. Cited on pages 42, 50, 93, and 109.

Carol Friedman. 2009. Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of the Electronic Health Record. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine: Artificial Intelligence in Medicine*, AIME '09, pages 1–5, Berlin, Heidelberg. Springer-Verlag. Cited on pages 34 and 35.

Jason Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, Matthias Samwald, and Wojciech Kusa. 2022. Dataset Debt in Biomedical Language Modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 137–145, virtual+Dublin. Association for Computational Linguistics. Cited on page 42.

William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102. Cited on page 87.

Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual Clinical NER: Translation or Cross-lingual Transfer? *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311. Cited on page 42.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92. Cited on page 128.

Oguzhan Gencoglu. 2020. Sentence Transformers and Bayesian Optimization for Adverse Drug Effect Detection from Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 161–164, Barcelona, Spain (Online). Association for Computational Linguistics. Cited on page 37.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, and Apurv Patki. 2014. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, pages 1–8. Cited on pages 34, 35, 38, and 39.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics. Cited on page 111.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. Cited on page 18.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*. Cited on page 32.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French Corpus with Clinical Cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics. Cited on page 110.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). Cited on page 29.

Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. Cited on page 24.

James Grimmelmann. 2017. The Law and Ethics of Experiments on Social Media Users. Preprint, LawArXiv. Cited on page 48.

Cyril Grouin, Natalia Grabar, Vincent Claveau, and Thierry Hamon. 2019. Clinical Case Reports for NLP. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 273–282, Florence, Italy. Association for Computational Linguistics. Cited on page 110.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karen Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*, pages 92–100. Cited on page 16.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23. Cited on page 37.

Harsha Gurulingappa, Abdul Mateen-Rajput, and Luca Toldo. 2012a. Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, 3:15. Cited on page 35.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012b. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892. Cited on pages 34, 35, 36, 37, and 38.

Andrey Gusev, Anna Kuznetsova, Anna Polyanskaya, and Egor Yatsishin. 2020. BERT Implementation for Detecting Adverse Drug Effects Mentions in Russian. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 46–50, Barcelona, Spain (Online). Association for Computational Linguistics. Cited on page 37.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical Named Entity Recognition with Multilingual BERT. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China. Association for Computational Linguistics. Cited on page 41.

Hasham Ul Haq, Veysel Kocaman, and David Talby. 2021. Deeper Clinical Document UnderstandingUsing Relation Extraction. *SDU 2022 workshop at AAAI*. Cited on page 37.

David Harbecke, Yuxuan Chen, Leonhard Hennig, and Christoph Alt. 2022. Why only Micro-F1? Class Weighting of Measures for Relation Classification. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 32–41, Dublin, Ireland. Association for Computational Linguistics. Cited on page 13.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H. Shah. 2014. Text Mining for Adverse Drug Events: The Promise, Challenges, and State of the Art. *Drug Safety*, 37(10):777–790. Cited on page 129.

Rave Harpaz, Krystl Haerian, Herbert S. Chase, and Carol Friedman. 2010. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annual Symposium Proceedings*, 2010:281–285. Cited on pages 34 and 35.

Zellig S. Harris. 1954. Distributional Structure. WORD, 10(2-3):146–162. Cited on page 21.

Lorna Hazell and Saad A. W. Shakir. 2006. Under-reporting of adverse drug reactions : A systematic review. *Drug Safety*, 29(5):385–396. Cited on page 1.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. Cited on page 25.

Leonhard Hennig, Philippe Thomas, and Sebastian Möller. 2023. MultiTACRED: A Multilingual Version of the TAC Relation Extraction Dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada. Association for Computational Linguistics. Cited on pages 31 and 32.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12. Cited on pages 42 and 109.

Andrew Herxheimer, Rose Crombag, and Teresa Leonardo Alves. 2010. Direct Patient Reporting of Adverse Drug Reactions, A Fifteen-Country Survey & Literature review. *Health Action International*. Cited on page 1.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023a. Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics. Cited on page 93.

Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2023b. Les textes cliniques français générés sont-ils dangereusement similaires à leur source ? Analyse par plongements de phrases. In *18e Conférence En Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs En RI – 30e Conférence Sur Le Traitement Automatique Des Langues Naturelles – 25e Rencontre Des Étudiants Chercheurs En Informatique Pour Le Traitement Automatique Des Langues*, pages 46–54, Paris, France. ATALA. Cited on page 93.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. Cited on page 21.

George Hripcsak and Daniel F. Heitjan. 2002. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2):99–110. Cited on pages 14 and 15.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12(3):296–298. Cited on pages 15 and 16.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR. Cited on pages 23, 30, 32, and 98.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 38.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse Drug Reaction Classification With Deep Neural Networks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 877–887. Cited on pages 34, 36, and 39.

Muhammad Ali Ibrahim, Muhammad Usman Ghani Khan, Faiza Mehmood, Muhammad Nabeel Asim, and Waqar Mahmood. 2021. GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification. *Journal of Biomedical Informatics*, 116:103699. Cited on page 33.

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. 2018. J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). Cited on page 86.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics. Cited on page 88.

Keyuan Jiang and Yujing Zheng. 2013. Mining Twitter Data for Potential Drug Effects. In *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 434–443, Berlin, Heidelberg. Springer. Cited on page 35.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035. Cited on page 36.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21. Cited on page 22.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020a. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. Cited on page 37.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics. Cited on page 23.

Dan Jurafsky and James H. Martin. 2023. Speech and Language Processing. 3rd (draft). Cited on pages 16, 18, 21, 22, 24, 26, 127, and 128.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81. Cited on pages 34, 45, 46, 47, 48, 50, 64, and 98.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1355–1360, Hong Kong, China. Association for Computational Linguistics. Cited on pages 30 and 32.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2014. Cross-Lingual Annotation Projection for Weakly-Supervised Relation Extraction. *ACM Transactions on Asian Language Information Processing*, 13(1):3:1–3:26. Cited on pages 30 and 31.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Internation Conference on Learning Representations*. Cited on page 106.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P Malek, Ulrich Keilholz, and Ulf Leser. 2021. Annotation and initial evaluation of a large annotated German oncological corpus. *JAMIA Open*, 4(2):ooab025. Cited on pages 42 and 109.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In *Fifth Social Media Mining for Health Applications(#SMM4H) Shared Tasks at COLING 2020*, page 10. Cited on pages 43, 45, 47, 48, and 61.

Ari Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, and Graciela Gonzalez. 2017. Detecting Personal Medication Intake in Twitter: An Annotated Corpus and Baseline Classification System. In *BioNLP 2017*, pages 136–142, Vancouver, Canada,. Association for Computational Linguistics. Cited on page 49.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics. Cited on page 45.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022. Alignment-Augmented Consistent Translation for Multilingual Open Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics. Cited on page 32.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: The Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956. Cited on page 41.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. SAGE. Cited on page 16.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–1079. Cited on pages 35 and 130.

Mayank Kulkarni, Daniel Preotiuc-Pietro, Karthik Radhakrishnan, Genta Indra Winata, Shijie Wu, Lingjue Xie, and Shaohua Yang. 2023. Towards a Unified Multi-Domain Multilingual Named Entity Recognition Model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2210–2219, Dubrovnik, Croatia. Association for Computational Linguistics. Cited on pages 30 and 32.

Vishesh Kumar, Amber Stubbs, Stanley Shaw, and Özlem Uzuner. 2015. Creation of a new longitudinal corpus of clinical narratives. *Journal of Biomedical Informatics*, 58 Suppl(Suppl):S6–S10. Cited on pages 105 and 110.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. (159). Cited on page 30.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics. Cited on pages 30 and 31.

Adam Lavertu and Russ B. Altman. 2019. RedMed: Extending drug lexicons for social media applications. *Journal of Biomedical Informatics*, 99:103307. Cited on page 49.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden. Association for Computational Linguistics. Cited on pages 34, 35, 42, and 44.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551. Cited on page 20.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323. Cited on page 20.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. Cited on pages 34, 36, and 37.

Kahyun Lee and Özlem Uzuner. 2020. Normalizing Adverse Events using Recurrent Neural Networks with Attention. *AMIA Summits on Translational Science Proceedings*, 2020:345–354. Cited on page 36.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55. Cited on page 82.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. Cited on page 93.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural Relation Extraction with Multilingual Attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43, Vancouver, Canada. Association for Computational Linguistics. Cited on pages 30 and 32.

D. A. Lindberg, B. L. Humphreys, and A. T. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291. Cited on page 129.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Cited on page 31.

H. P. Luhn. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4):309–317. Cited on page 22.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/UW shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, page ocad012. Cited on pages 33 and 79.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics. Cited on page 30.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192. Cited on pages 37, 38, and 103.

Diwakar Mahajan, Jennifer J. Liang, and Ching-Huei Tsou. 2021. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2021:833–842. Cited on pages 105, 106, 108, and 110.

Chris Manning. 2006. Doing Named Entity Recognition? Don't optimize for F1. Cited on page 13.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schuetze. 2009. *Introduction to Information Retrieval*. Cambridge UP. Cited on page 4.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap Translation for Cross-Lingual Named Entity Recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics. Cited on page 30.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 591–598. Morgan Kaufmann. Cited on page 30.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press. Cited on page 99.

Rebecca McKee. 2013. Ethical issues in using social media for health and health care research. *Health Policy*, 110(2):298–301. Cited on page 50.

Charles Medawar, Andrew Herxheimer, Andrew Bell, and Shelley Jofre. 2002. Paroxetine, Panorama and user reporting of ADRs: Consumer intelligence matters in clinical practice and post-marketing drug surveillance. *International Journal of Risk & Safety in Medicine*, 15(3):161–169. Cited on page 1.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large Language Models as Instructors: A Study on Multilingual Clinical Entity Extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada. Association for Computational Linguistics. Cited on page 42.

Alejandro Metke-Jimenez, Sarvnaz Karimi, and Cecile Paris. 2014. Evaluation of text-processing algorithms for adverse drug event extraction from social media. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, SoMeRA '14, pages 15–20, New York, NY, USA. Association for Computing Machinery. Cited on pages 34, 35, 38, 39, 45, 46, and 48.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina. 2020. KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56, Barcelona, Spain (Online). Association for Computational Linguistics. Cited on pages 37 and 41.

Jude Mikal, Samantha Hurst, and Mike Conway. 2016. Ethical issues in using Twitter for population-level depression monitoring: A qualitative study. *BMC Medical Ethics*, 17(1):22. Cited on pages 50 and 51.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*. Cited on pages 22, 23, 29, and 128.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. Cited on pages 29 and 30.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Red Hook, NY, USA. Curran Associates Inc. Cited on pages 22, 23, and 29.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics. Cited on page 18.

Shubhanshu Mishra and Aria Haghighi. 2021. Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 381–388, Online. Association for Computational Linguistics. Cited on pages 31 and 32.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery. Cited on page 128.

Tom M. Mitchell. 1997. *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill, New York. Cited on page 16.

Megan A. Moreno, Alison Grant, Lauren Kacvinsky, Peter Moreno, and Michael Fleming. 2012. Older Adolescents' Views Regarding Participation in Facebook Research. *Journal of Adolescent Health*, 51(5):439–444. Cited on page 51.

Stephen Mutuvi, Emanuela Boros, Antoine Doucet, Gaël Lejeune, Adam Jatowt, and Moses Odeo. 2020. Multilingual Epidemiological Text Classification: A Comparative Study. In *COLING, International Conference on Computational Linguistics*, page 6172. Cited on page 41.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26. Cited on page 11.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2021. A Data Bootstrapping Recipe for Low-Resource Multilingual Relation Classification. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 575–587, Online. Association for Computational Linguistics. Cited on page 32.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2):10:1–10:21. Cited on page 92.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12. Cited on page 40.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics. Cited on page 92.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The Quaero French Medical Corpus: A ressource for medical entity recognition and normalization. page 7. Cited on page 110.

Aurélie Névéol, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. 2015. CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical named entity recognition. *CLEF (Working Notes)*. Cited on page 14.

Aurelie Neveol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Gregoire Rey, and Pierre Zweigenbaum. 2018. CLEF eHealth 2018 Multilingual Information Extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian. In *Conference and Labs of the Evaluation Forum*. Cited on page 41.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The Scielo Corpus: A Parallel Corpus of Scientific Publications for Biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA). Cited on page 41.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics. Cited on pages 30, 31, and 32.

Jian Ni and Radu Florian. 2019. Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 399–409, Hong Kong, China. Association for Computational Linguistics. Cited on page 30.

Azadeh Nikfarjam and Graciela H. Gonzalez. 2011. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2011:1019–1026. Cited on pages 34 and 35.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015a. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association: JAMIA*, 22(3):671–681. Cited on page 35.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015b. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681. Cited on page 49.

Tomohiro Nishiyama, Mihiro Nishidani, Aki Ando, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. NAISTSOC at the NTCIR-16 Real-MedNLP Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo. Cited on page 86.

Naoaki Okazaki. 2007. CRFsuite: A fast implementation of Conditional Random Fields (CRFs). Cited on page 36.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on page 42.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. Cited on pages 40 and 42.

Caterina Palleria, Christian Leporini, Serafina Chimirri, Giuseppina Marrazzo, Sabrina Sacchetta, Lucrezia Bruno, Rosaria M. Lista, Orietta Staltari, Antonio Scuteri, Francesca Scicchitano, and Emilio Russo. 2013. Limitations and obstacles of the spontaneous adverse drugs reactions reporting: Two "challenging" case reports. *Journal of Pharmacology & Pharmacotherapeutics*, 4(Suppl1):S66–S72. Cited on pages 1 and 2.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. Cited on page 19.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics. Cited on pages 30 and 32.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured Prediction as Translation between Augmented Natural Languages. In *International Conference on Learning Representations*. Cited on page 38.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. Cited on page 93.

Apurv Patki, Abeed Sarker, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen Oconnor, Karen Smith, and Graciela Gonzalez. 2014. Mining Adverse Drug Reaction Signals from Social Media: Going Beyond Extraction. Cited on pages 34 and 35.

Michael J. Paul, Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith, and Graciela Gonzalez. 2016. SOCIAL MEDIA MINING FOR PUBLIC HEALTH MONITORING AND SURVEILLANCE. In *Biocomputing 2016*, pages 468–479, Kohala Coast, Hawaii, USA. WORLD SCIENTIFIC. Cited on pages 49 and 50.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830. Cited on pages 13 and 131.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. Cited on pages 22, 23, and 29.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *35th Conference on Neural Information Processing Systems*, page 17. Cited on page 99.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. Cited on page 29.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics. Cited on page 32.

Daniel Pimienta. 2022. Resource: Indicators on the Presence of Languages in Internet. In *Proceedings of SIGUL2022*, pages 83–91. Cited on page 23.

Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. 2021. BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1740–1747, Online. Association for Computational Linguistics. Cited on page 37.

David M. W. Powers. 2020. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Cited on page 13.

Marko Prelevikj and Slavko Zitnik. 2021. Multilingual Named Entity Recognition and Matching Using BERT and Dedupe for Slavic Languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 80–85, Kiyv, Ukraine. Association for Computational Linguistics. Cited on page 32.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959. Cited on page 22.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551. Cited on pages 37, 38, 42, 50, 53, and 86.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*. Cited on page 12.

Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Yue Yu, and Hongfang Liu. 2016. Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*. Cited on page 35.

Shivam Raval, Hooman Sedghamiz, Enrico Santus, Tuka Alhanai, Mohammad Ghassemi, and Emmanuele Chersoni. 2021. Exploring a Unified Sequence-To-Sequence Transformer for Medical Product Safety Monitoring in Social Media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3534–3546, Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on pages 37, 38, and 41.

Dietrich Rebholz-Schuhmann, Simon Clematide, Fabio Rinaldi, Senay Kafkas, Erik M. van Mulligen, Chinh Bui, Johannes Hellrich, Ian Lewin, David Milward, Michael Poprat, Antonio Jimeno-Yepes, Udo Hahn, and Jan Kors. 2013. Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. *Lecture Notes in Computer Science*, pages 353–367. Cited on page 40.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. Cited on pages 37 and 93.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia. Association for Computational Linguistics. Cited on page 30.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. 2020. Soft Gazetteers for Low-Resource Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online. Association for Computational Linguistics. Cited on page 30.

C. J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworths. Cited on page 13.

Raul Rodriguez-Esteban. 2009. Biomedical Text Mining and Its Applications. *PLOS Computational Biology*, 5(12):e1000597. Cited on page 33.

Roland Roller, Ammer Ayach, and **Lisa Raithel**. 2021. Boosting transformers using background knowledge, or how to detect drug mentions in social media using limited data. In *Proceedings of the BioCreative VII Challenge Evaluation Workshop*. Not cited.

Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke, and Bilgin Osmanodja. 2022. An Annotated Corpus of Textual Explanations for Clinical Decision Support. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2317–2326, Marseille, France. European Language Resources Association. Cited on page 110.

Roland Roller, Madeleine Kittner, Dirk Weissenborn, and Ulf Leser. 2018. Cross-lingual Candidate Search for Biomedical Concept Normalization. *11th Language Resources and Evaluation Conference*. Cited on page 41.

Sebastian Ruder. 2016. On word embeddings - Part 1. Cited on pages 22 and 23.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway. Cited on pages 19 and 20.

Sebastian Ruder. 2022. The State of Multilingual AI. Cited on page 23.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533. Cited on pages 18 and 21.

Noboru Sakai. 2013. The role of sentence closing as an emotional marker: A case of Japanese mobile phone e-mail. *Discourse, Context & Media*, 2(3):149–155. Cited on page 91.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *EMC2̂: 5th Edition Co-located with NeurIPS'19*. Cited on page 37.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019b. A Hierarchical Multi-Task Approach for Learning Embeddings from Semantic Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6949–6956. Cited on page 33.

Sunita Sarawagi. 2008. Information Extraction. *Foundations and Trends in Databases*, 1(3):261–377. Cited on page 3.

Abeed Sarker. 2017. Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. Cited on page 49.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M. Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. 2018. Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1274–1283. Cited on page 34.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212. Cited on page 43.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207. Cited on pages 34, 35, 38, 39, and 49.

Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2016. Social Media Mining Shared Task Workshop. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 21:581–592. Cited on pages 34 and 36.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513. Cited on page 36.

Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2021. NADE: A Benchmark for Robust Adverse Drug Events Extraction in Face of Negations. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 230–237, Online. Association for Computational Linguistics. Cited on page 38.

Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2022. Increasing adverse drug events extraction robustness on social media: Case study on negation and speculation. *Experimental Biology and Medicine (Maywood, N.J.)*, page 15353702221128577. Cited on pages 2, 3, 35, and 38.

Sanja Scepanovic, Enrique Martin-Lopez, Daniele Quercia, and Khan Baykaner. 2020. Extracting medical entities from social media. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 170–181, New York, NY, USA. Association for Computing Machinery. Cited on page 49.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62, Seattle, WA. Association for Computational Linguistics. Cited on page 42.

H. J. A. Schouten. 1980. Measuring pairwise agreement among many observers. *Biometrical Journal*, 22(6):497–504. Cited on pages 46 and 47.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. Cited on page 26.

H. Schütze. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, Supercomputing '92, pages 787–796, Washington, DC, USA. IEEE Computer Society Press. Cited on page 22.

Holger Schwenk and Matthijs Douze. 2017. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics. Cited on pages 30 and 88.

Holger Schwenk and Xian Li. 2018. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). Cited on pages 30 and 31.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19:321–325. Cited on page 16.

Isabel Segura-Bedmar, Ricardo Revert, and Paloma Martínez. 2014. Detecting drugs and adverse events from Spanish social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 106–115, Gothenburg, Sweden. Association for Computational Linguistics. Cited on pages 1, 2, 35, 39, 43, 46, 47, 48, 50, and 71.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. Cited on page 31.

Jurica Seva, Mario Sänger, and Ulf Leser. 2018. WBI at CLEF eHealth 2018 Task 1: Language-independent ICD-10 Coding using Multi-lingual Embeddings and Recurrent Neural Networks. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*. Cited on page 41.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting Anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics. Cited on page 33.

Kanaka D. Shetty and Siddhartha R. Dalal. 2011. Using information mining of the medical literature to improve drug safety. *Journal of the American Medical Informatics Association: JAMIA*, 18(5):668–674. Cited on page 34.

Stefano Silvestri, Francesco Gargiulo, Mario Ciampi, and Giuseppe De Pietro. 2020. Exploit Multilingual Language Model at Scale for ICD-10 Clinical Text Classification. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7. Cited on page 33.

Chenchen Song. 2022. Sentence-final particle vs. sentence-final emoji: The syntax-pragmatics interface in the era of CMC. Cited on page 91.

Stefania Spina. 2019. Role of Emoticons as Structural Markers in Twitter Interactions. *Discourse Processes*, 56(4):345–362. Cited on page 91.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015a. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, 58 Suppl(Suppl):S11–S19. Cited on pages 105 and 110.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015b. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of biomedical informatics*, 58 Suppl(Suppl). Cited on pages 105 and 110.

Corey Sutphin, Kahyun Lee, Antonio Jimeno Yepes, Özlem Uzuner, and Bridget T. McInnes. 2020. Adverse drug event detection using reason assignments in FDA drug labels. *Journal of Biomedical Informatics*, 110:103552. Cited on page 36.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics. Cited on page 120.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics. Cited on page 92.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. DAMO-NLP at SemEval-2023 Task 2: A Unified Retrieval-augmented System for Multilingual Named Entity Recognition. Cited on page 30.

Ludovic Tanguy, Lydia-Mai Ho-Dac, Cécile Fabre, Roxane Bois, Touati Mohamed Yacine Haddad, Claire Ibarboure, Marie Joyau, François Le moal, Jade Moiilic, Laura Roudaut, Mathilde Simounet, Irena Stankovic, and Mickaela Vandewaetere. 2020. LITL at SMM4H: An Old-school Feature-based Classifier for Identifying Adverse Effects in Tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 134–137, Barcelona, Spain (Online). Association for Computational Linguistics. Cited on page 37.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433. Cited on page 27.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 30.

Lisa Raithel, Philippe Thomas, Roland Roller, Oliver Sapina, Sebastian Möller, and Pierre Zweigenbaum. 2022. Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3637–3649, Marseille. European Language Resources Association. Cited on pages 6, 59, and 97.

Lisa Raithel*, Hui-Syuan Yeh*, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Philippe Thomas, Sebastian Möller, Tomohiro Nishiyama, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum. 2024. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages. In *Proceedings of the Language Resources and Evaluation Conference*, Torino. European Language Resources Association. Not cited.

Philippe Thomas, Mariana Neves, Illes Solt, Domonkos Tikk, and Ulf Leser. 2011. Relation Extraction for Drug-Drug Interactions using Ensemble Learning. *DDIExtraction2011: First Challenge Task: Drug-Drug Interaction Extraction at SEPLN 2011*. Cited on page 34.

Paul Thompson, Sophia Daikou, Kenju Ueno, Riza Batista-Navarro, Jun'ichi Tsujii, and Sophia Ananiadou. 2018. Annotation and detection of drug effects in text for pharmacovigilance. *Journal of Cheminformatics*, 10(1):37. Cited on pages 34 and 36.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature. *PLOS Computational Biology*, 6(7):e1000837. Cited on page 34.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. Cited on page 75.

D. Yu. Turdakov, N. A. Astrakhantsev, Ya. R. Nedumov, A. A. Sysoev, I. A. Andrianov, V. D. Mayorov, D. G. Fedorenko, A. V. Korshunov, and S. D. Kuznetsov. 2014. Texterra: A framework for text analysis. *Programming and Computer Software*, 40(5):288–295. Cited on page 44.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2020. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews. *arXiv:2004.03659 [cs]*. Cited on pages 37, 43, 44, 45, 46, 47, and 48.

Özlem Uzuner, Amber Stubbs, and Leslie Lenert. 2020. Advancing the state of the art in automatic extraction of adverse drug events from narratives. *Journal of the American Medical Informatics Association*, 27(1):1–2. Cited on pages 36 and 39.

Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. 2012. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884. Cited on page 35.

Vladimir Naumovich Vapnik and Alexey Yakovlevich Chervonenkis. 1964. A note on one class of perceptrons. *Automation and Remote Control*, 25(1):112–120. Cited on page 20.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017*, page 11. Cited on pages 24, 25, 26, 29, and 32.

Harsh Verma, Sabine Bergler, and Narjesossadat Tahaei. 2023. Comparing and combining some popular NER approaches on Biomedical tasks. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 273–279, Toronto, Canada. Association for Computational Linguistics. Cited on page 40.

Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zoë Thomas, and John-Paul Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database: The Journal of Biological Databases and Curation*, 2013:bat019. Cited on page 14.

Andreas Vilhelmsson. 2015. Consumer Narratives in ADR Reporting: An Important Aspect of Public Health? Experiences from Reports to a Swedish Consumer Organization. *Frontiers in Public Health*, 3:211. Cited on page 1.

Shoko Wakamiya, Mizuki Morita, and Yoshinobu Kano. 2017. Overview of the NTCIR-13: MedWeb Task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo. Cited on page 86.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. Cited on page 27.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018b. Adversarial Multilingual Neural Relation Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1156–1166, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Cited on pages 30 and 32.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending Multilingual BERT to Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics. Cited on page 32.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics. Cited on pages 34, 36, 37, and 49.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16, Brussels, Belgium. Association for Computational Linguistics. Cited on pages 36 and 49.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association. Cited on page 31.

P. J. Werbos. 1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Ph.D. thesis, Harvard University. Cited on page 21.

P.J. Werbos. 1990. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560. Cited on page 21.

Karin Wester, Anna K. Jönsson, Olav Spigset, Henrik Druid, and Staffan Hägg. 2008. Incidence of fatal adverse drug reactions: A population based study. *British Journal of Clinical Pharmacology*, 65(4):573–579. Cited on page 1.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv:1910.03771 [cs]*. Cited on pages 106 and 131.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics. Cited on page 128.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason

Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Cited on pages 26 and 98.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics. Cited on pages 30, 31, and 32.

Shuntaro Yada, Shoko Wakamiya, Yuta Nakamura, and Eiji Aramaki. 2022. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies,*. Cited on page 86.

Christopher C. Yang, Haodong Yang, Ling Jiang, and Mi Zhang. 2012. Social media mining for drug safety signal detection. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, SHB '12, pages 33–40, New York, NY, USA. Association for Computing Machinery. Cited on pages 1 and 2.

Ming Yang, Xiaodi Wang, and Melody Kiang. 2013. Identification of Consumer Adverse Drug Reaction Messages on Social Media. *PACIS 2013 Proceedings*. Cited on page 35.

Xi Yang, Jiang Bian, Ruogu Fang, Ragnhildur I. Bjarnadottir, William R. Hogan, and Yonghui Wu. 2020. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *Journal of the American Medical Informatics Association: JAMIA*, 27(1):65–72. Cited on page 36.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on page 33.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual Seq2seq Training with Similarity Loss for Cross-Lingual Document Classification. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 175–179, Melbourne, Australia. Association for Computational Linguistics. Cited on page 30.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. Cited on page 93.

Minghao Zhu and Keyuan Jiang. 2021. Semi-Supervised Language Models for Identification of Personal Health Experiential from Twitter Data: A Case for Medication Effects. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 228–237, Online. Association for Computational Linguistics. Cited on page 37.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27. Cited on page 128.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu,

Soo Kyung Park, Kelly Xu, Hamideh Moayyed, and Somaieh Goudarzvand. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *Journal of Biomedical Informatics*, 90:103091. Cited on pages 1, 3, 36, 45, 46, 47, 48, 50, 61, and 98.

Maryam Zolnoori, Timothy Patrick, Kin Fung, Paul Fontelo, Anthony Faiola, Shirley Wu, Kelly Xu, Jiaxi Zhu, and Christina Eldredge. 2017. Development of an Adverse Drug Reaction Corpus from Consumer Health Posts. In *SMM4H@AMIA*, Washington, DC, USA. Cited on page 47.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial Feature Adaptation for Cross-lingual Relation Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Cited on page 32.

Résumé

Le travail décrit dans cette thèse porte sur la détection et l'extraction d'effets indésirables de médicaments dans des textes biomédicaux rédigés par le grand public, c'est-à-dire par des personnes qui ne sont pas des professionnels de santé, à travers les barrières linguistiques et multilingues.

Les effets indésirables de médicaments sont des réactions nuisibles ou désagréables résultant de l'utilisation d'un produit médical et peuvent parfois être mortelles. Des erreurs de dosage, l'automédication, des diagnostics incorrects, des allergies ou d'autres conditions non détectées du patient, des abus, ainsi que des interactions avec d'autres médicaments ou substances peuvent provoquer ces réactions. De fait, les effets indésirables de médicaments constituent un problème de santé majeur dans le monde entier principalement en raison du grand nombre d'effets indésirables qui passent inaperçus.

En premier lieu, cela est dû aux essais cliniques, qui ne sont généralement pas menés sur des groupes vulnérables spécifiques (comme les personnes âgées ou enceintes). Deuxièmement, il est tout simplement impossible de représenter toute une population. Bien qu'il existe divers mécanismes de signalement et de post-surveillance, ces derniers sont souvent méconnus des patients et des professionnels de santé, le signalement est chronophage et il manque de détails. De plus, même s'ils sont signalés, les rapports ne reflètent généralement que la perspective du médecin, et rarement celle du patient. Enfin, la langue joue également un rôle important dans les soins de santé au quotidien. Les personnes qui ne parlent pas parfaitement la ou les langues officielles d'un pays sont souvent désavantagées lorsqu'elles communiquent avec leur médecin. Cela réduit encore davantage les chances de signaler des effets indésirables de médicaments.

De ce fait, les médias sociaux constituent une ressource précieuse pour rassembler des connaissances concernant les effets secondaires indésirables. De nos jours, de nombreuses personnes ont accès à Internet et utilisent les médias sociaux. Elles fournissent ainsi des informations dans différentes langues et avec différents contextes sociaux et éthiques. Cela nous permet d'observer des gens qui s'expriment « à leur manière » et, surtout, de manière anonyme. Ce dernier point est crucial lorsque des effets indésirables surviennent, par exemple, en lien avec la sexualité. De plus, la langue dans laquelle le grand public parle de problèmes médicaux nous permet de comprendre ces problèmes de santé différemment de si l'on se fiait uniquement aux rapports des médecins, qui utilisent généralement une langue plus technique. Cela rend par ailleurs ces expériences plus accessibles et susceptibles d'être partagées avec d'autres patients.

Les sources souvent utilisées en ce qui concerne les médias sociaux sont Twitter, Reddit ou des groupes Facebook. Les forums de patients offrent une alternative, souvent déjà axée sur des discussions orientées vers la santé, parfois même administrées par des experts. En tenant compte de tout ce qui précède, ce travail vise à améliorer la manière dont les connaissances sur les effets indésirables des médicaments peuvent être rassemblées et traitées, à travers les langues et du point de vue des patients, en utilisant des messages sur les médias sociaux dans différentes langues, principalement issus de forums de patients. Ce travail est principalement réalisé pour les langues allemande, française et anglaise, mais comprend également quelques expérimentations utilisant des textes en japonais et en espagnol.

Tout d'abord, j'aborde le contexte technique nécessaire pour comprendre les concepts appliqués par la suite. Cela inclut la définition des tâches d'extraction d'informations qui sont

prises en compte, à savoir la classification binaire, l'extraction d'entités et de relations, ainsi que leur évaluation. Je donne également un aperçu des méthodes d'évaluation de l'annotation de corpus. Ensuite, je passe en revue les concepts sous-jacents à l'apprentissage automatique et discute brièvement du transfert d'apprentissage. Ceci est suivi d'un résumé des systèmes et des modèles couramment utilisés pour les tâches mentionnées ci-dessus, ainsi que d'une brève introduction à la modélisation de la langue multilingue et translingue, en mettant un accent particulier sur les modèles basés sur les Transformers.

Dans le chapitre suivant, je passe en revue les travaux connexes dans le domaine de l'extraction d'informations (multilingue). Je commence par me concentrer sur l'extraction d'informations translingue et multilingue en général, mettant en évidence les façons courantes de définir la tâche d'extraction et les modèles sous-jacents pour aborder ces tâches.

Le chapitre se poursuit ensuite avec un aperçu de l'extraction d'informations spécifiquement dans le domaine biomédical. Les modèles et les jeux de données sont examinés, et les méthodes standard sont présentées, divisées en celles de la période pré-apprentissage automatique, de l'apprentissage automatique traditionnel et de l'apprentissage profond, y compris les approches basées sur les Transformers. La section se termine sur un aperçu des défis souvent rencontrés dans le domaine biomédical.

La troisième section de ce chapitre traite de l'extraction d'informations translingues dans le domaine biomédical. Encore une fois, les jeux de données, les modèles et les méthodes sont présentés.

Ensuite, je donne un aperçu détaillé des ensembles de données pour la détection des effets secondaires médicaux dans des langues autres que l'anglais, ainsi que des différences dans leurs annotations. Deux ensembles de données en anglais, qui sont utilisés dans ce travail, sont également discutés.

Enfin, le chapitre se conclut par un bref aperçu des problématiques de respect de la vie privée des utilisateurs lors de la collecte de données. La section aborde les différentes méthodes de collecte de données et les types de données souvent utilisés dans le traitement de textes biomédicaux, et résume les problèmes éthiques qui les accompagnent.

La partie la plus significative de cette thèse est consacrée aux données, c'est-à-dire à leur collecte, annotation et analyse. Je commence le quatrième chapitre en décrivant le nouveau corpus fourni dans cette thèse. Le corpus est disponible en allemand et en français et fait partie d'un ensemble de données trilingue, qui comprend également des données japonaises. Je discute de l'élaboration de directives d'annotation applicables à n'importe quelle langue et axées sur les textes créés par les utilisateurs de médias sociaux. En outre, je présente le processus de collecte de données, ainsi que les critères pour traduire une partie des données allemandes en français. Cela est suivi par la description de l'annotation binaire du corpus allemand. Ce corpus se compose d'environ 10 000 documents avec des étiquettes binaires, où l'étiquette positive représente la présence d'un effet indésirable médical, tandis que l'étiquette négative marque les documents sans mention d'effet indésirable. Les scores d'accord inter-annotateurs et les statistiques de l'ensemble de données sont présentés. Je fournis également l'accord des annotateurs avec les données adjudiquées, montrant des détails intéressants du processus d'annotation. En conclusion, je constate que le corpus créé est très exigeant, tant en termes d'annotation que pour les modèles d'apprentissage automatique, car les données sont souvent ambiguës, les documents peuvent être très longs et il y a un déséquilibre très élevé des étiquettes, avec seulement 324 documents positifs au total.

L'annotation se poursuit sur une partie plus petite du corpus mentionné précédemment, c'est-à-dire uniquement les documents positifs (pour le moment). Ici, les entités, attributs et relations sont annotés pour les données françaises et allemandes. Je décris les directives et le schéma d'annotation pour 12 types d'entités, quatre types d'attributs et 12 types de relations (plus quelques autres pour une analyse interne). Pour capturer les expressions médicales, des

marqueurs pour les médicaments, les signes, les examens médicaux, les mentions anatomiques, les expressions temporelles et les opinions personnelles des patients sont fournis, entre autres. Les mentions de médicaments peuvent être complétées par des attributs d'état de médicament, indiquant si un médicament vient d'être démarré, arrêté, augmenté ou diminué. Les opinions peuvent avoir un sentiment positif, négatif ou neutre, et les signes et fonctions peuvent être niés. De plus, les expressions temporelles peuvent être marquées avec des balises plus spécifiques indiquant si la mention fait référence à une durée ou à un moment précis. Enfin, les types d'entités peuvent être associés par l'utilisation de relations. Elles indiquent, par exemple, quel résultat a eu un examen médical ou quelle forme pharmaceutique a été utilisée pour un médicament spécifique. La relation la plus importante est celle de « cause » entre un médicament et un signe, qui marque les effets indésirables potentiels des médicaments, la combinaison qui nous intéresse le plus. En tout, pour l'allemand, 118 documents sont soigneusement annotés et vérifiés. Pour le français, un annotateur a actuellement annoté 100 documents. Les deux corpus comprennent respectivement 3 487 et 1 939 entités, 1 141 et 537 attributs, et 2 163 et 1 129 relations. Les annotations continuent.

Ensuite, je discute de la question de la vie privée des utilisateurs en ce qui concerne les données liées à la santé, ainsi que de la manière de collecter de telles données à des fins de recherche sans nuire à la vie privée de la personne. Je présente une étude prototype sur la réaction des utilisateurs lorsqu'on leur demande directement leurs expériences de réactions indésirables aux médicaments. Elle est basée sur une enquête que j'ai diffusée sur plusieurs plateformes, notamment Reddit et deux plates-formes d'enquête. Dans le but de recueillir des descriptions écrites d'expériences avec les effets indésirables, les participants sont d'abord invités à consentir à partager leurs expériences personnelles à des fins de recherche. Ensuite, on leur pose plusieurs questions démographiques auxquelles ils peuvent choisir de ne pas répondre. On leur demande ensuite d'indiquer leurs médicaments et leur diagnostic (s'ils existent) sous la forme qui leur convient, et enfin, s'ils ont connu des effets secondaires. Pour ces derniers, on leur demande de les décrire aussi en détail que possible, sans utiliser de puces. Le questionnaire a été distribué en allemand et en anglais. Au final, 54 participants ont répondu aux questions, mais seuls 27 ont terminé. Leurs réponses montrent une grande variété de descriptions concernant les médicaments, les dosages et les effets secondaires. En effet, les participants se sont montrés assez ouverts avec leurs textes, fournissant des documents pas très longs, mais néanmoins complets, similaires aux avis en ligne sur les médicaments. En résumé, l'étude révèle que la plupart des gens n'ont pas d'objection à décrire leurs expériences s'ils en sont directement sollicités. Cependant, la collecte de données peut souffrir d'un questionnaire comportant trop de questions.

Dans la section suivante, j'analyse une deuxième façon potentielle de collecter des données, à savoir la génération synthétique de données basée sur de vrais messages Twitter, et les défis que cela pose. Ce travail visait spécifiquement la diffusion potentielle de données, ce qui est plus compliqué pour les tweets originaux en raison des préoccupations susmentionnées concernant la vie privée. Les tweets ont été générés en japonais, annotés pour les signes et symptômes médicaux, puis automatiquement traduits en anglais, en français et en allemand. Les étiquettes annotées sont reprises. Malgré les tentatives de filtrer et d'améliorer les valeurs aberrantes dans les pseudo-tweets traduits, je constate encore des problèmes dans les traductions, tant en ce qui concerne le sens du texte que les étiquettes annotées. Ces pseudo-tweets sont ensuite analysés plus en détail dans cette thèse, et je donne des exemples anecdotiques de ce qui peut mal tourner lors de la traduction automatique. Par exemple, les traductions ne sont pas toujours cohérentes d'une langue à l'autre et de légères variations changent le sens. De plus, elles affichent souvent des inexactitudes médicales, ce qui n'est pas nécessairement un problème pour les besoins de ce travail, mais doit être pris en compte. Elles montrent également souvent des biais de différents types qui pourraient influencer la performance et la généralisabilité d'un modèle lorsqu'il est affiné sur ces données. À la fin de la section, je résume les

enseignements tirés et présente les étapes potentielles pour améliorer davantage le corpus.

Ensuite, je résume les expériences réalisées sur le transfert translingue de connaissances concernant les effets indésirables de médicaments en anglais et en allemand, c'est-à-dire la classification binaire de documents contenant (ou non) des mentions de réactions indésirables. Une partie du corpus allemand annoté de manière binaire est utilisée pour évaluer les performances interlingues, ce qui est compliqué par le déséquilibre important des ensembles de données des deux langues. Les expériences sont menées en tenant compte d'une configuration à ressources limitées, ce qui était en effet le cas pour les données allemandes. Bien qu'il y ait environ 4 000 documents à traiter, seuls 101 d'entre eux étaient positifs et ils devaient également être répartis entre les ensembles d'entraînement, de développement et de test. J'ai donc mené les expériences en deux étapes. La première étape consistait en un affinage sur les données anglaises uniquement, sur des données contenant des effets indésirables de médicaments mais avec une distribution d'étiquettes inversée, c'est-à-dire plus de documents positifs que négatifs. Ensuite, j'ai appliqué, dans la deuxième étape, plusieurs scénarios avec des tailles de jeux de données et des ratios d'étiquettes variables. Cela incluait l'équilibrage des documents par étiquette de classe, l'ajout d'une certaine quantité d'échantillons négatifs aux positifs (c'est-à-dire le contrôle des négatifs, mais en utilisant tous les positifs), et l'ajout d'une quantité spécifique d'échantillons anglais aux données d'affinage. Enfin, j'ai utilisé toutes les données d'entraînement allemandes disponibles pour l'affinage, c'est-à-dire un grand nombre d'exemples négatifs et un très faible nombre d'exemples positifs. J'ai comparé les résultats de ces expériences à un modèle de base utilisant un séparateur à vaste marge et à l'application d'un modèle sans entraînement spécifique à l'allemand.

Les résultats de ces différentes approches ont démontré qu'incorporer des données d'entraînement anglaises aide à détecter des documents pertinents en allemand. Cependant, cela ne suffit pas à compenser le déséquilibre naturel des étiquettes des documents allemands. Le meilleur modèle était en effet celui qui n'avait été affiné qu'avec toutes les données d'entraînement allemandes, en utilisant les données telles quelles, sans sur-échantillonnage ni sous-échantillonnage. Cependant, bien que le score F1 global pour la classe positive de la plupart des modèles soit très faible, les scores de rappel étaient souvent suffisamment élevés pour que l'on puisse envisager d'utiliser ces modèles comme filtres pour rassembler plus de données pertinentes, ce qui, à son tour, conduirait très probablement à de meilleures performances pour la classe positive.

Dans le sixième chapitre, je décris d'abord ma participation à la campagne d'évaluation n2c2 2022 concernant la détection de médicaments. À cette fin, les organisateurs ont fourni un jeu de données constitué de textes de dossiers de patients en anglais, annotés avec des mentions de médicaments (et d'autres annotations). Pour les expériences sur la détection de médicaments, j'ai utilisé plusieurs modèles basés sur les Transformers cliniques/biomédicaux, affiné plusieurs modèles de chaque type, et combiné les prédictions résultantes. Les résultats obtenus avec cette approche étaient plutôt bons, mais présentaient encore certaines faiblesses. Tout d'abord, le nombre de faux positifs surpassait le nombre de faux négatifs, c'est-à-dire que le modèle généralisait trop, en particulier sur les termes médicaux qui n'étaient pas des mentions de médicaments, mais étaient utilisés dans des contextes similaires. De plus, nous avons trouvé des incohérences dans les annotations et des expressions devenant (en apparence) similaires à des noms de médicaments en raison de fautes de frappe. Les mentions que les modèles ont manquées étaient principalement des abréviations et des traitements ambigus.

Sur la base des conclusions de la campagne d'évaluation, les expériences ont ensuite été étendues à d'autres langues, à savoir le français, l'allemand et l'espagnol, en utilisant des jeux de données appartenant à différents sous-domaines et basés sur des directives d'annotation différentes. Pour ces langues, trois jeux de données en allemand, deux en français, deux en espagnol et le jeu de données déjà utilisé en anglais ont été collectés et préparés. J'ai ensuite

affiné un modèle multilingue du domaine général sur différents sous-ensembles de langues pour voir quelle combinaison est la plus bénéfique pour chaque langue. J'ai comparé cela à un modèle affiné sur toutes les langues. Les expériences menées montrent que le transfert multilingue et interlingue fonctionne, mais qu'aucune méthode ne donne des scores aussi élevés que ceux pour les données en anglais. J'ai constaté que les performances du modèle dépendent fortement des types d'annotations et de leurs définitions ainsi que de la structure du texte, qui est très variable d'un jeu de données à l'autre. Dans l'analyse des erreurs, j'ai de nouveau trouvé plus de faux positifs que de faux négatifs. Comme précédemment, il s'agit souvent d'incohérences dans les annotations ou de médicaments spécifiques qui ne sont pas annotés dans un corpus, mais annotés dans un autre. Cela concerne souvent les noms de médicaments qui sont très similaires d'une langue à l'autre.

Sur la base du corpus précédemment annoté de messages de patients annotés avec des mentions de médicaments et d'autres entités, j'ai appliqué les modèles entraînés décrits ci-dessus aux nouvelles données sans réentraînement (en mode *zero-shot*) afin d'obtenir quelques résultats préliminaires. Le modèle affiné uniquement en allemand (et non dans les autres langues) a obtenu les meilleurs résultats sur la partie allemande du corpus. En revanche, le modèle entraîné dans toutes les langues a obtenu les meilleurs scores sur les données françaises. Les modèles sont améliorables, mais ils fournissent déjà une première base prometteuse, surtout compte tenu du fait que les modèles ont été affinés sur des données d'un autre sous-domaine et appliqués sans réentraînement aux nouvelles données qui contiennent de nombreuses expressions courantes non standard.

Je conclus le chapitre sur quelques expériences préliminaires pour la détection générale d'entités dans le nouveau corpus en français et en allemand. J'ai constaté que les résultats varient considérablement entre les types d'entités, ce qui n'est pas surprenant étant donné les différentes exigences de chaque type d'entité.

Dans le dernier chapitre, je résume le travail présenté et le replace dans le contexte d'autres travaux dans ce domaine. De plus, en m'appuyant sur les résultats et les questions subséquentes qui émergent des différents chapitres, je propose des idées pour étendre davantage les recherches sur la détection et la prévention des effets indésirables de médicaments à travers différentes langues.