Development of Artificial Intelligence Methods for the Analysis of Online Data for Medical Research Purposes : Use Case on the World Diabetes Distress Study

Adrian Ahne

université PARIS-SACLAY

Development of artificial intelligence methods for the analysis of online data for medical research purposes: Use case on the World Diabetes Distress Study

Développement de méthodes d'intelligence artificielle pour l'analyse de données de réseaux sociaux et à des fins de recherche médicale : Cas d'utilisation sur une étude mondiale sur le diabète

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570 : santé publique (EDSP)
Spécialité de doctorat : santé publique - épidémiologie
Graduate School : Santé Publique. Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche **CESP (Université Paris-Saclay, UVSQ, Inserm)**, sous la direction de **Guy FAGHERAZZI**, Directeur de recherche, Luxembourg Institute of Health et la co-supervision de **Thomas CZERNICHOW**, directeur e-santé Epiconcept

Thèse soutenue à Paris-Saclay, le 25 janvier 2022, par

Adrian AHNE

Composition du Jury

Pascale TUBERT-BITTER Directrice de recherche, INSERM, université Paris Saclay	Présidente
Sandra BRINGAY Professeure, université Montpellier	Rapportrice & Examinatrice
Gayo DIALLO Maître de conférence, HDR, université Bordeaux	Rapporteur & Examinateur
Marie-Aline CHARLES Directrice de recherche, INSERM, université Sorbonne	Examinatrice
Adam HULMAN Senior Data Scientist, Steno Diabetes Center Aarhus	Examinateur
Guy FAGHERAZZI Directeur de recherche, Luxembourg Institute of Health	Directeur de thèse

THÈSE DE DOCTORAT
NNT : 2022UPASR001

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Guy Fagherazzi, for accepting me first as an intern and then PhD student, for the trust he placed in me, for the freedom he gave me and the constant support in work but also non-work related matters. Thank you for opening me the door to the research world, outstanding supervision, including quick answers to mails within hours, and those enriching 3 years in which I learnt a lot!

A sincere thank you as well to Epiconcept, who supported me financially and allowed me to pursue this thesis project. In particular, my tutors Dr. Thomas Czernichow and Francisco Orchard who were always there for valuable advice, fruitful discussion, brainstorming around machine learning and in particular debugging *FeedbackExplorer*.

An additional thank you goes to my thesis committee: Prof. Dr. Sandra Bringay and Dr. Gayo Diallo for having accepted the roles of *rapporteurs* in my jury, Dr. Marie-Aline Charles and Dr. Adam Hulman for having accepted the roles as *examiners* and Dr. Pascale Tubert-Bitter for having accepted the role as the president for my thesis committee.

I would like to express my appreciation to Prof. Xavier Tannier, who accompanied me from my studies at Polytech Sorbonne to my PhD project, for his helpful advice in various NLP related questions and his subtle comments in our publications which significantly improved the papers.

A big thank you to Vivek Khetan, who supported and advised me a lot during our causality work and our long discussions about life in the US and Europe despite the great time difference that separated us between San Francisco and Paris, and which gave rise to amusing discrepancies; he got up at dawns whereas I was about to go to bed.

A special thank you to all my colleagues from the E3N/E4N team, their warm welcome and cheerfulness. In particular to Marco, Joe, Conor, Douae, Thibault, Fanny, Pauline and Yahya with whom I found many new friendships and countless discussions about life, politics and Swing dance in cozy atmosphere of bars. The presence of these people certainly made the thesis time more joyful. Besides, I would also like to thank the current director of E3N/E4N Dr. Gianluca for his open ear and fascinating political exchanges.

Thank you to my dearest friends, for their continuous encouragement and support with a special mention to Bilge for his help in managing D3 when I got stuck.
Finally, this thesis would not have been possible without the support of my family, notably the small messages of encouragement of my sister. Particularly, I would like to thank my mother, who unknowingly helped in the feasibility of this thesis by trusting my intuition to study abroad and making it possible. The latest thanks goes to my partner in life Anne who made me not only critically reflect on my work during these three years but also for her presence and unconditional support.

SCIENTIFIC PRODUCTION

ARTICLES ACCEPTED FOR PUBLICATION

Ahne A, Orchard F, Tannier X, Perchoux C, Balkau B, Pagoto S, Harding JL, Czernichow T, Fagherazzi G. *Insulin pricing and other major diabetes-related concerns in the USA: a study of 46 407 tweets between 2017 and 2019*. BMJ Open Diabetes Research and Care 2020;8:e001190. doi: 10.1136/bmjdrc-2020-001190

Ahne A, Fagherazzi G, Tannier X, Czernichow T, Orchard F. *Improving diabetes-related biomedical literature exploration in the clinical decision-making process via Interactive classification and topic discovery: Methodology Development Study*. Journal of Medical Internet Research 2022;24(1):e27434 doi: 10.2196/27434

ARTICLES SUBMITTED

Ahne A, Khetan V, Tannier X, Rizvi MIH, Czernichow T, Orchard F, Bour C, Fano A, Fagherazzi G. *Identifying causal relationships in tweets using deep learning: Use case on diabetes-related tweets from 2017-2021*. 2021
Preprint available on arXiv: https://arxiv.org/abs/2111.01225

OTHER PUBLICATIONS

Fagherazzi G, **Ahne A**, Guillot C, Riveline JP, Bonnet F, Mebarki A, Schuck S, Czernichow T, Jeannerod G, Orchard F. *Étude mondiale de la détresse liée au diabète : le potentiel du réseau social Twitter pour la recherche médicale*. Revue d'Épidémiologie et de Santé Publique, Volume 66, Supplement 4, 2018, Pages S197-S198, ISSN 0398-7620, doi: https://doi.org/10.1016/j.respe.2018.04.002.

Bour C, Schmitz S, **Ahne A**, Perchoux C, Dessenne C, Fagherazzi G. *Scoping review protocol on the use of social media for health research purposes.* BMJ Open 2021;11:e040671. doi: 10.1136/bmjopen-2020-040671

Bour C, **Ahne A,** Schmitz S, Perchoux C, Dessenne C, Fagherazzi G. *The Use of Social Media for Health Research Purposes: Scoping Review.* J Med Internet Res 2021;23(5):e25736. DOI: 10.2196/25736

SCIENTIFIC COMMUNICATIONS

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *Identification of diabetes distress patterns based on social media data using artificial intelligence methods: the world diabetes distress study*.
International Diabetes Epidemiology Group Conference, Seoul, 2019, Oral presentation

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *Identification of diabetes distress patterns based on social media data using artificial intelligence methods: the world diabetes distress study*.
European Diabetes Epidemiology Group Conference, Luxembourg, 2019, Oral presentation

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *Identification of diabetes distress patterns based on social media data using artificial intelligence methods: the world diabetes distress study*.
Data Science Summer School Conference, Paris, École Polytechnique, 2019, Poster presentation

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *Memory efficient and targeted topic discovery in medical documents*
AI4 Health School conference - Paris, Health Data Hub, PR[AI]RIE, MiAi Grenoble Alpes, 3iA Côte d'Azur, 2021, Poster presentation

OTHER COMMUNICATIONS

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *World Diabetes Distress Study: To better understand diabetes distress, using Twitter and AI methods*.
Luxembourg Institute of Health, Deep Digital Phenotyping Research Unit, 2020, Online presentation

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *Identification of diabetes distress patterns based on social media data using artificial intelligence methods: the world diabetes distress study*.
Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS), Inserm, 2020

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *Improving the evidence-based clinical decision making process: Interactive classification and topic discovery on diabetes-related biomedical literature*
Machine Learning Journal Club, Steno Diabetes Center Aarhus, Denmark, 2021, Online presentation

Ahne A, Orchard F, Czernichow T, Fagherazzi G. *Identification of diabetes distress patterns based on social media data using artificial intelligence methods: the world diabetes distress study*.
Global Diabetes Journal Club (GDJC), 2021, Online presentation

Ahne A, Khetan V, Tannier X, Rivzi Md IH, Czernichow T, Orchard F, Bour C, Fano A, Fagherazzi G. *Identification of cause and associated effect events in diabetes-related Tweets*.
Accenture Labs, San Francisco, January 2022, Online presentation

TRAINING

Animation of an atelier on *Causality extraction in diabetes-related Tweets.*
November 2021, Datacraft, Sorbonne Center for Artificial Intelligence

AWARD

Price in winning the 3 minutes thesis competition (ma thèse en 180 secondes)

International Diabetes Epidemiology Group Conference - Seoul - 2019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ANNEXES

LIST OF ABBREVIATIONS

AL: Active learning

BERT: Bidirectional Encoder Representation from Transformer

CDSS: Clinical decision support system

DD: Diabetes Distress

DOC: Diabetes Online Community

EBM: Evidence based medicine

EHR: Electronic Health Record

FBE: Feedback Explorer

FCLL: Fully connected linear layers

GDM: Gestational Diabetes Mellitus

HAC: Hierarchical agglomerative clustering

HbA_{1c}: Glycated Hemoglobin

HC: Hierarchical clustering

ML: Machine learning

NER: Named entity recognition

NLP: Natural language processing

OGTT: Oral glucose tolerance test

SM: Social media

SVM: Support Vector Machine

T1D: Type 1 Diabetes

T2D: Type 2 Diabetes

RÉSUMÉ EN FRANÇAIS

CHAPITRE I: INTRODUCTION

L'intelligence artificielle transforme le secteur de la santé à une vitesse fulgurante en intervenant dans tous les secteurs. Parallèlement, avec l'adoption massive et l'intégration des données en ligne et réseaux sociaux dans nos vies, une nouvelle source de données a émergé qui peut être exploitée à des fins épidémiologiques.

Le diabète est une maladie chronique qui touche 463 millions d'adultes (âgés de 20-79 ans) dans le monde en 2019 et devrait atteindre 700 millions en 2045.[1]
Cette thèse de doctorat explore les méthodes d'intelligence artificielle pour l'analyse des données des médias sociaux et le développement d'un système d'aide à la décision clinique. Le cas d'usage de ce travail est le diabète et la détresse liée au diabète dans le cadre de la World Diabetes Distress Study.

1.1 Diabète

1.1.1 Types de diabète

Le diabète peut être classé en trois types principaux. Le diabète de type 1 (DT1) est causé par une réponse auto-immune dans laquelle le corps attaque les cellules bêta productrices d'insuline, conduisant à une production d'insuline insuffisante ou inexistante.[1] Les origines de cette réponse auto-immune ne sont pas entièrement connues, mais les éléments environnementaux et génétiques sont suspectés.[2] Jusqu'à aujourd'hui, le seul traitement existant est l'injection quotidienne d'insuline pour maintenir une glycémie stable. Le diabète de type 1 affecte le plus souvent les patients dans l'enfance, mais les symptômes peuvent parfois se développer plus tard et représentent 5 à 10 % de tous les cas de diabète.[1,3] Dans le diabète de type 2 (DT2), le corps perd sa capacité à répondre correctement à l'insuline, également appelée résistance à l'insuline.[4] Les origines de la résistance à l'insuline ne sont pas entièrement comprises, mais les facteurs de risque importants sont le surpoids, l'obésité, l'âge, l'origine ethnique et les antécédents familiaux.[1,5] C'est

le type de diabète le plus courant concernant environ 90 % de tous les cas de diabète dans le monde.[1,6] Enfin, le diabète gestationnel est défini comme une hyperglycémie qui se développe d'abord au cours de la grossesse et sa prévalence varie fortement selon les régions (Europe : 6,1 %, Asie du Sud-Est : 15 %, Amérique du Nord : 7 %).[1,7]

1.1.2 Épidémie de diabète

La prévalence mondiale du diabète chez les adultes âgés de 20 à 79 ans était estimée à 463 millions en 2019 (9,3 % de la population mondiale totale dans ce groupe d'âge) et devrait augmenter régulièrement pour atteindre 700 millions en 2045 (10,9 % de la population mondiale).[1]

1.1.3 Détresse liée au diabète

Un sous-concept central du diabète est la détresse liée au diabète (DD) qui regroupe des facteurs psychologiques, tels que le stress, les inquiétudes, les émotions, la fatigue et l'impuissance liés à la gestion quotidienne du diabète.[8,9] Tout au long de cette thèse, nous nous référons également aux profils de détresse liés au diabète, qui englobent les problèmes quotidiens liés au diabète exprimés par les personnes atteintes de diabète. Les deux mesures les plus courantes et validées de DD sont les questionnaires auto-déclarés: Problem Areas in Diabetes (PAID) et Diabetes Distress Scale (DDS).[10,11] Par contre, nous avons encore des connaissances limitées sur toutes les sources de préoccupations, de stress et d'anxiété chez les personnes atteintes de diabète, y compris celles émergentes, et les échelles actuelles ne capturent pas l'image complète de la DD. Les limites de ces échelles incluent également l'auto-déclaration entraînant des inexactitudes potentielles dans les données.[12]

Une méta-analyse en 2017 a signalé une prévalence globale de 36 % pour la détresse liée au diabète chez les personnes atteintes de diabète de type 2 et une prévalence significativement plus élevée chez les femmes.[13] Chez les personnes atteintes de diabète de type 1, la prévalence de la DD est estimée à environ 20 à 40 %.[9,14]

1.2 Réseaux sociaux

1.2.1 Qu'est-ce que c'est?

Le terme réseaux sociaux est apparu au début des années 2000 avec la disponibilité croissante de l'accès Internet haut débit et peut être défini comme des applications Web dont la fonction principale est le développement et l'échange de contenu généré par les utilisateurs.[15,16] Les réseaux

sociaux offrent un moyen direct et facile à utiliser d'échanger et de réseauter avec d'autres personnes par rapport aux médias traditionnels.

En avril 2021, 4,7 milliards de personnes correspondant à 60% de la population mondiale utilisaient Internet.[17] Parmi ces internautes, 4,3 milliards étaient des utilisateurs actifs des réseaux sociaux (55% de la population mondiale) et passaient en moyenne 2 heures 22 minutes par jour sur les réseaux sociaux, montrant la popularité et l'omniprésence des réseaux sociaux dans toutes nos vies.[17] Ils influencent la façon dont nous vivons, travaillons et communiquons les uns avec les autres.

Aux États-Unis, la plateforme en ligne la plus populaire est Youtube avec 81 % d'utilisateurs, suivie de Facebook avec 69 % et d'Instagram (40 %).[18] Twitter se situe au milieu de terrain avec 23% d'utilisateurs actifs et une plus grande popularité parmi les 18-29 ans par rapport aux personnes plus âgées. Les plus jeunes préfèrent Snapchat, Instagram, Facebook et YouTube, tandis que les personnes plus âgées sont généralement moins présentes sur les autres médias sociaux que Facebook et YouTube.[19]

1.2.2 Le cas de Twitter

La thèse est centrée sur les données de la plateforme Twitter, une plateforme de micro-blogging avec 396 millions d'utilisateurs actifs en avril 2021 et 500 millions de tweets envoyés chaque jour.[17,20] Une particularité de Twitter par rapport à ses homologues des médias sociaux est son caractère public. Généralement, les informations partagées sur Twitter sont publiques et potentiellement consultables par tous dans le monde, y compris par les personnes sans compte Twitter.[21]

La majorité des utilisateurs de Twitter sont des hommes (68,5 %) et âgés de moins de 50 ans, 25,2 % des utilisateurs ayant entre 18 et 24 ans, 26,6 % entre 25 et 34 ans et 28,4 % entre 35 et 49 ans.[17] Une enquête sur les utilisateurs adultes de Twitter aux États-Unis à partir de 2019 a montré leur niveau d'éducation élevé, avec 42% ayant obtenu un diplôme universitaire, contre 31% dans la population générale des États-Unis.[22] En outre, les utilisateurs de Twitter se situent souvent de manière disproportionnée dans la classe des revenus élevés avec 41% gagnant plus de 75 000 $, contre 32% de la population générale qui dépasse ce seuil.[22]

1.2.3 Les médias sociaux pour la recherche en santé publique

Le caractère public de Twitter et par conséquent l'accès facile aux données est certainement la principale raison de son attrait pour les chercheurs.[23] Une enquête a confirmé que les utilisateurs des réseaux sociaux échangent sur leurs expériences personnelles de santé avec des amis ou en groupe, collectent des fonds pour attirer l'attention sur un problème de santé spécifique ou recherchent des informations sur la santé.[24]
Le suivi de ces informations, traitements et sentiments liés à la santé, liés aux publications ou aux discussions sur les réseaux sociaux, offre de nouvelles opportunités de développer des méthodes pour améliorer les soins de santé.[25–27]

Bour et al. ont fourni un aperçu détaillé de l'utilisation des médias sociaux pour la recherche en santé.[28] Ils ont observé le caractère évolutif de l'utilisation des médias sociaux dans la recherche en santé, de la concentration sur les maladies transmissibles (par exemple la grippe, le VIH)[29,30] dans des études antérieures à l'inclusion et plus récemment, des stratégies de recrutement et de la collecte de données pour l'infoveillance (surveillance syndromique par internet).[28,31]

Cependant, la mauvaise qualité et la fiabilité des médias sociaux représentent des obstacles à leur facilité d'utilisation.[32] Contrairement à la médecine fondée sur les faits, qui ne met pas l'accent sur les histoires anecdotiques, les médias sociaux ont tendance à les souligner.[33] Les atteintes à la vie privée des patients peuvent potentiellement causer beaucoup plus de dommages sur les réseaux sociaux qu'en face à face avec le professionnel de la santé en raison du caractère ouvert du SM et de la permanence des informations numériques.[34] D'autres problèmes éthiques liés au SM à des fins de santé sont l'obtention du consentement des utilisateurs en ligne, en particulier pour les adolescents ou leurs parents, et la préservation de l'anonymat des participants à l'étude.[35] Pour le moment, aucune directive générale officielle n'existe pour traiter ces questions éthiques sur les médias sociaux.[35]

1.3.4 Réseaux sociaux pour la recherche sur le diabète

Les médias sociaux, tels que Twitter, sont de plus en plus populaires parmi les personnes atteintes de diabète.[36] Dans une analyse qualitative de diverses plateformes SM, les chercheurs ont montré que les personnes atteintes de diabète partagent des conseils pratiques, se connectent en groupe et partagent des éléments de leur quotidien avec humour et fierté.[37] Une autre revue souligne le rôle bénéfique des communautés en ligne sur le diabète avec relativement peu de conséquences

négatives.[38] Potentiellement, ces plateformes offrent une forme de soutien social aux personnes atteintes de diabète, ce qui semble être le principal moteur de l'adhésion au traitement.[39] Béguerisse-Diaz et al. ont trouvé des groupes thématiques tels que les informations sur la santé, les actualités, les interactions sociales, les tweets commerciaux et humoristiques.[40]

Les interventions sur les réseaux sociaux, par le biais d'appareils technologiques (par exemple, les téléphones portables) pour les programmes d'éducation et de sensibilisation à la santé, ont un impact bénéfique sur la réduction de l'HbA1c et augmentent la satisfaction des patients.[41,42] De plus, il a été démontré que les réseaux sociaux ont un effet positif sur la réduction de l'HbA1c et l'autogestion.[43,44]

Le crowdsourcing pour recueillir des informations auprès de la communauté en ligne du diabète est également une application utile, par exemple, pour la surveillance de la sécurité des dispositifs pour le diabète.[45,46]

1.3 Langage naturel dans l'aide à la décision clinique

La quantité exponentiellement croissante de données textuelles en ligne avec un intérêt clinique (dossiers de santé électroniques, interactions médecin-patient, réseaux sociaux, etc.) rend difficile pour les professionnels de santé d'analyser et d'extraire des informations précieuses et, par conséquent, d'exercer une médecine basée sur les faits. Dans le système de santé actuel, les professionnels de la santé sont fréquemment confrontés à des situations dans lesquelles ils ne savent pas que des données spécifiques sur les patients sont accessibles en ligne, ni comment y accéder, n'ont pas le temps de les rechercher ou ne sont pas à jour avec les dernières avancées en matière de recherche médicale.[47,48]

Des outils intelligents d'aide à la décision clinique permettent de sortir de ce dilemme. Ils permettent de filtrer efficacement ces données, d'extraire les informations pertinentes et de les présenter de manière compréhensible. Cependant, les freins qui entravent l'adoption à grande échelle de ces outils dans la pratique clinique sont des interfaces utilisateur médiocres, l'incapacité à intégrer ces outils naturellement dans la routine du professionnel de la santé, la conception de ces outils d'un point de vue trop technique, ou un manque d'acceptation globale par les médecins.[49–51] Dans un rapport, Bates et al. ont partagé dix enseignements tirés de la mise en œuvre de systèmes

d'aide à la décision clinique, dans lesquels ils soulignent l'importance de la convivialité de l'outil, permettant aux professionnels de la santé de « faire ce qui est juste ».[52]

Une autre raison importante de l'adoption limitée est la nature « black box » des algorithmes d'apprentissage automatique incorporés dans ces systèmes, et leur manque d'interprétabilité pour les médecins et les ingénieurs qui les ont développés.[53,54] En réponse à ce phénomène, une feuille de route sur l'aide à la décision clinique a été publiée par l'American Medical Informatics Association (AMIA) identifiant trois directions principales pour parvenir à une large adoption : 1) Les meilleures connaissances disponibles en cas de besoin; 2) Adoption élevée et utilisation efficace; 3) Amélioration continue des connaissances et méthodes des systèmes d'aide à la décision clinique.[55]

Dans cette thèse, nous visons à examiner la première direction à suivre pour structurer, analyser et présenter au mieux les connaissances et comment les lacunes connues entravant l'adoption de systèmes d'aide à la décision clinique peuvent être comblées, telles que l'interprétabilité, la facilité d'utilisation et la gestion de grands corpus de données.

CHAPITRE II: OBJECTIFS

Aborder une question de recherche en épidémiologie traditionnelle repose principalement sur des hypothèses. Contrairement à cette stratégie, dans cette thèse, une approche axée sur les données est suivie. Les apports de ce travail sont doubles. Sur le plan épidémiologique, cette thèse exploratoire vise à extraire des données liées au diabète des informations qui pourraient ensuite être utilisées dans une prochaine étape pour formuler des hypothèses qui sont étudiées dans un cadre plus traditionnel. Côté technique, l'objectif était le développement et la validation d'algorithmes d'apprentissage automatique innovants, qui ont ensuite été rendus open-source pour inciter la communauté à les réutiliser pour leurs travaux de recherche.

2.1 Profils de diabète et de détresse liés au diabète

Le premier objectif de cette thèse était d'identifier les profils de diabète et de détresse liés au diabète et les préoccupations liées au diabète à l'aide des données des réseaux sociaux. À notre connaissance, il s'agissait de la première étude à recueillir des informations sur les principales préoccupations liées au diabète à l'aide des données des réseaux sociaux.

2.2 Relations causales

Actuellement, il est difficile d'identifier automatiquement les associations de cause à effet dans les données textuelles. Dans le deuxième objectif de cette thèse, la détection de paires cause-effet dans les tweets liés au diabète a été étudiée en utilisant de puissants algorithmes d'apprentissage automatique. À l'instar du premier objectif, il s'agit de la première étude portant sur l'identification des paires cause-effet dans les tweets liés au diabète.

2.3 *FeedbackExplorer* - Aide à la décision clinique

Les professionnels de la santé pratiquent la médecine factuelle sur la base des meilleures connaissances disponibles. La quantité exponentiellement croissante de données biomédicales (littérature, DSE, réseaux sociaux) pose de sérieuses difficultés pour exercer correctement l'EBM sur la base d'informations récentes et à jour. Le troisième objectif aborde la partie synthèse de la littérature dans le processus de prise de décision clinique et traite du développement d'une méthodologie pour structurer, analyser et visualiser efficacement la littérature scientifique en mettant l'accent sur l'interprétabilité, la facilité d'utilisation pour les praticiens de la santé sans connaissances en programmation et la capacité à gérer de grands corpus de données.

CHAPITRE III: CONCEPTS GÉNÉRAUX

3.1 Données

La principale source de données pour le premier et le deuxième objectifs était Twitter. Les données ont été téléchargées via l'interface de programmation d'application de Twitter permettant aux utilisateurs de définir des mots-clés liés au diabète et de renvoyer les tweets correspondants à partir d'un échantillon aléatoire de 1 % de tous les tweets. Toutes les données recueillies dans ce travail ont été publiées publiquement sur Twitter. Conformément à la politique de confidentialité de Twitter, les utilisateurs s'engagent à rendre ces informations accessibles au grand public.[21]

Le troisième objectif évalue l'outil de prise de décision clinique sur sa capacité à structurer les résumés PubMed, une grande base de données bibliographiques biomédicales contenant plus de 33 millions de références à des articles de revues biomédicales et des sciences de la vie à partir de

1946 et fournies par la National Library of Medicine.[56] Les résumés PubMed sont classés hiérarchiquement via des codes MeSH (Medical Subject Headings), utilisés pour l'indexation, le catalogage et la recherche de documents biomédicaux.

3.2 Apprentissage automatique et traitement du langage naturel

L'apprentissage automatique est une branche de l'intelligence artificielle conçue pour apprendre de l'environnement et de l'expérience.[57] Une fois l'apprentissage d'un algorithme terminé, il est capable de généraliser et de traiter des données invisibles par rapport à la tâche souhaitée.

Une architecture spécifique des algorithmes d'apprentissage automatique sont les réseaux de neurones (NN), ou les réseaux de neurones artificiels (ANN), qui ont été inspirés par les efforts visant à simuler le fonctionnement du cerveau avec les neurones, étant les nœuds, et les connexions entre eux.[58] Dans les réseaux multicouches ou à action directe, les neurones sont disposés en couches dans lesquelles plusieurs couches cachées sont placées entre la couche d'entrée et la couche de sortie. Une architecture de réseau de neurones comportant plusieurs couches cachées empilées les unes sur les autres est communément appelée apprentissage en profondeur (“Deep learning”).[59]

Le langage humain est complexe, rempli d'ambiguïtés et sujet à l'interprétation d'une personne, ce qui rend extrêmement difficile le développement de programmes qui déterminent de manière fiable le sens voulu des données textuelles. Le traitement du langage naturel (NLP) explore comment les ordinateurs peuvent être utilisés pour comprendre et manipuler le langage naturel.[60]

3.3 Représentation des données

Les algorithmes d'apprentissage automatique fonctionnent généralement avec des données numériques et préfèrent des entrées et des sorties de longueur fixe bien définies, alors que les données textuelles sont généralement désordonnées. Ainsi, les données textuelles, telles que les données des réseaux sociaux, doivent être transformées en nombres. Deux percées clés ont entraîné l'essor du traitement du langage naturel ces dernières années et ont révolutionné la façon dont les données textuelles sont transformées en nombres utiles : premièrement, la capacité à générer des représentations vectorielles significatives de mots, appelées *word embeddings* ou vecteurs denses, déclenchée par la méthode *word2vec*[61,62] et étendu par *FastText*;[63] et deuxièmement, d'améliorer les méthodes mentionnées précédemment par la possibilité de générer des représentations de mots et de phrases sensibles au contexte[64] à l'aide d'une architecture de type “Transformer”.[65]

Les *word embeddings FastText* constituent la base des trois objectifs et le modèle *BERT* joue un rôle crucial dans notre deuxième objectif d'identifier les associations causales.

Tout au long de ce travail, la “similarité cosinus” a été appliquée pour déterminer la similarité entre les mots, les phrases ou les tweets.

3.4 Algorithmes supervisés

Les algorithmes d'apprentissage automatique concernent la reconnaissance de formes dans les données. Ils sont un outil utile pour distinguer les données pertinentes des données non pertinentes. Les méthodes d'apprentissage supervisée apprennent un modèle à partir de données d'entraînement étiquetées pour effectuer des prédictions sur des données invisibles.

L'algorithme d'apprentissage automatique supervisé Support vector machine (SVM)[66,67] vise à trouver un hyperplan optimal qui sépare au mieux les classes en maximisant la distance entre l’hyperplan lui-même et les points de données les plus proches. Dans le premier et le deuxième objectif de cette thèse, le SVM a été utilisé dans le pipeline de prétraitement pour filtrer uniquement les tweets pertinents et, par conséquent, supprimer le bruit.

3.5 Algorithmes non supervisés

Les algorithmes non supervisés visent à trouver des regroupements de documents similaires dans une collection de documents.

L'algorithme de clustering K-means est un algorithme non supervisé simple et largement utilisé qui crée une décomposition « plate » des données en un ensemble de clusters.[68] Le premier objectif a utilisé l'algorithme K-means pour détecter les clusters à partir de tweets.

Le clustering hiérarchique (HC) est une forme de clustering dans laquelle la solution est présentée sous forme d'arbres. Le clustering hiérarchique est idéalement situé pour l'exploration et la visualisation interactives, car il permet l'extraction de diverses partitions plates de solutions de clustering de différents niveaux de granularité.[69] Un nouvel algorithme de regroupement hiérarchique a été développé dans le troisième objectif, jetant les bases de l'outil d'aide à la décision clinique.

3.6 Reconnaissance d'entité nommée

La reconnaissance des entités nommées (NER) vise à identifier les entités appropriées et à les classer dans des catégories prédéfinies telles que les dates, les produits, les noms ou dans le domaine biomédical des maladies, des facteurs de risque, des protéines ou des mutations. Généralement, la tâche de NER est formalisée comme une tâche de classification de séquences. Un modèle d'étiquetage de séquence est formé pour attribuer une classe d'étiquette à chaque jeton ou unité au sein d'une séquence de jetons.

Une méthode standard pour identifier les entités dans un texte sont les champs aléatoires conditionnels (CRF), une méthode de classification séquentielle statistique,[70] qui a été appliquée dans le deuxième objectif pour détecter les associations causales.

3.7 Apprentissage actif

L'annotation manuelle des données est essentielle pour le développement et l'évaluation réussis de systèmes d'apprentissage automatique sophistiqués. En même temps, cela prend du temps, est très coûteux et reste donc un défi pour les groupes de recherche.[71] L'apprentissage actif (AL) est une méthode de sélection d'échantillons visant à minimiser les coûts d'annotation, tout en maximisant les performances des modèles d'apprentissage automatique, en sélectionnant les données d'entraînement de manière intelligente.[72]
La stratégie d'échantillonnage d'incertitude est la plus simple et la plus couramment utilisée, dans laquelle l'instance est choisie pour laquelle le système est le moins sûr de la façon d'étiqueter.[73]
Dans le deuxième et le troisième objectif, une stratégie d'apprentissage actif, basée sur un échantillonnage d'incertitude, a été adoptée pour augmenter efficacement les données d'entraînement et augmenter les performances tout en minimisant l'effort d'annotation.

3.8 Métriques d'évaluation

Les tâches de classification binaire étaient généralement évaluées sur la base des valeurs: les vrais positifs (TP) sont définis comme le nombre d'instances correctement identifiées et étiquetées comme appartenant à la classe positive. Les vrais négatifs (TN) correspondent aux instances correctement identifiées et étiquetées comme appartenant à la classe négative. Les faux positifs (FP) font référence au nombre d'instances incorrectement identifiées appartenant à la classe positive et les faux négatifs (FN) spécifient le nombre d'instances incorrectement identifiées comme appartenant à la classe négative.

Les mesures de performance typiques tout au long de cette thèse étaient: $accuracy = TP + TN / (TP + FP + FN + TN)$; $precision = TP / (TP + FP)$; $recall = TP / (TP + FN)$ and $F1 = 2 + (precision * recall) / (precision + recall)$.

Ces mesures peuvent également être étendues pour les tâches de classifications multi-classes.

CHAPITRE IV: IDENTIFICATION DES PROFILS DE DÉTRESSE DIABÈTE

4.1 Méthodologie

Un pipeline de prétraitement strict a été appliqué pour supprimer les tweets bruyants et non pertinents. Les retweets et les doublons ont été supprimés. Les SVM ont été exploitées pour s'entraîner sur des données étiquetées manuellement afin de se concentrer sur les tweets avec un contenu personnel sans blague. Un moteur de géolocalisation a été développé pour déduire l'emplacement d'un tweet en fonction de ses métadonnées. Dans une dernière étape de prétraitement, seuls les tweets contenant un élément émotionnel (mot émotionnel ou emoji/émoticônes) ont été conservés, pour se concentrer sur la détresse liée au diabète, les facteurs psychologiques et les émotions. Les mots émotionnels ont été identifiés sur la base du travail de Parrot qui nous a fourni une liste de plus de 100 émotions classées dans une structure hiérarchique avec six émotions primaires : joie, amour, surprise, tristesse, colère et peur. Une analyse des sentiments a été menée pour inspecter l'opinion exprimée par les utilisateurs dans leurs communications textuelles, si l'opinion ou l'attitude a tendance à être positive ou négative en adoptant l'outil open source et validé par le “Valence Aware Dictionary for Sentiment Reasoning” (VADER).[74]

Le cœur de ce travail était l'identification des profils de diabète et de détresse liée au diabète et de caractériser ce dont parlent les personnes atteintes ou liées au diabète. Ceci a été réalisé en utilisant un algorithme de Kmeans. Le paramètre d'entrée du nombre de clusters a été estimé à l'aide du score Silhouette et a conduit à un nombre optimal de 30 clusters.

Une extension de ce travail a été l’étude du croisement des sujets d'intérêt identifiés avec la variable socio-économique du revenu moyen des ménages basée sur l'enquête communautaire américaine 2017 du US Census Bureau.[75] Le revenu moyen des ménages a été divisé en tertiles : faible revenu (US$ 24.609 - US$ 67.224); revenu moyen (moyen) (US$ 67.225 - US$ 86.758) et revenu élevé (US$ 86.759 - US$ 394.259). Chaque tweet a ensuite été lié au revenu moyen de sa ville géolocalisée et affecté à son tertile respectif de revenu faible, moyen ou élevé, ce qui nous a permis de calculer les associations entre les thèmes et les tertiles de revenu moyen.

En outre, le sexe et le type de diabète ont été estimés à l'aide de SVM sur des données étiquetées manuellement.

4.2 Résultats

Dans ce travail, nous avons montré que Twitter est un outil utile pour capturer et décrire les principaux sujets liés au diabète et les émotions liées à ces clusters. Nos principales conclusions suggèrent que d'une part, parmi la communauté en ligne du diabète, il existe beaucoup de soutien mutuel et de solidarité avec de multiples tweets contenant des éléments de joie et d'amour. D'autre part, que les utilisateurs partageaient des émotions de peur, de colère et de tristesse en ce qui concerne le prix de l'insuline et les complications et comorbidités liées au diabète. En outre, l'incapacité des gens à faire la distinction entre les différents types de diabète a causé une grande frustration. Alors que les femmes sur-représentent la plupart des sujets, en particulier les sujets se référant à l'importance d'une insuline abordable et du test de tolérance au glucose par voie orale (OGTT), les hommes ont tendance à discuter de sujets autour d'histoires liées au diabète et au "choc diabétique/insuline". De plus, nous avons constaté que dans les villes à revenus plus élevés, on était plus susceptibles de discuter de sujets concernant l'insuline abordable, le prix de l'insuline, l'utilisation de la langue liée au diabète ou à l'instabilité glycémique. Alors que les discussions dans les villes à faible revenu moyen étaient davantage axées sur des histoires quotidiennes liées au diabète, les échanges dans la communauté en ligne (DSMA) et le test oral de tolérance au glucose.

4.3 Discussion

Une préoccupation majeure parmi les tweets aux États-Unis était le prix de l'insuline (sujets sur 5/30) avec tout le spectre des émotions présentes. Les émotions positives (joie, amour) se sont manifestées dans des tweets faisant référence à la solidarité dans la lutte pour une insuline abordable au sein de la communauté du diabète. Des émotions négatives (tristesse, colère, peur) étaient présentes lorsque les gens partageaient leur frustration concernant les prix de l'insuline, l'accès à l'insuline et l'identification des sources d'insuline, y compris les "gardiens du glucose" ou les dons, qui constituent des obstacles majeurs pour les personnes atteintes de diabète.[76,77]

Le premier avantage crucial de l'utilisation des données des réseaux sociaux est le fait que les informations sont exprimées spontanément et en temps réel. Cela peut être décrit comme un espace numérique ouvert avec une absence de hiérarchie dans les rôles pour le partage d'informations et le développement de communautés en ligne. En conséquence, les biais potentiels survenant dans les études traditionnelles et observationnelles, tels que le biais du répondant, pourraient être minimisés, car il n'y a pas de hiérarchie entre les parties. Deuxièmement, un grand

nombre de personnes et une grande variabilité dans leurs profils ont été analysés. Troisièmement, nous avons développé une méthodologie innovante pour nous concentrer sur des tweets géolocalisés pertinents (personnels, émotionnels, sans blague) en provenance des États-Unis afin d'identifier les sujets d'intérêt et les émotions partagées au sein des sujets. Enfin, cette approche est capable de capturer les tendances de la communauté en ligne du diabète et les facteurs socio-économiques qui peuvent être associés à un niveau écologique.

La principale limite est que les utilisateurs qui expriment des préoccupations liées au diabète sur Twitter peuvent ne pas être représentatifs de toutes les personnes atteintes de diabète. Néanmoins, il a été suggéré que cela peut être partiellement compensé par le grand nombre et malgré tout la diversité des personnes partageant des données en premier lieu, une force majeure de l'épidémiologie numérique.[27] Malgré l'observation d'une grande variabilité dans les profils, nous avons repéré une surreprésentation des personnes atteintes de diabète de type 1 et des femmes par rapport à la littérature épidémiologique du diabète connue alors qu'en réalité la grande majorité des cas de diabète sont de type 2 (90 %).[1] Une deuxième limite tient à ce que les performances, en particulier la précision, de nos classifieurs n'étaient pas parfaitement précises, ce qui signifie qu'il n'y a aucune garantie que tous nos tweets aient véritablement été publiés par des personnes atteintes de diabète partageant du contenu personnel et il n'a pas toujours été possible de déterminer le sexe ou le type de diabète. Troisièmement, nous n'avons pas tenu compte des facteurs cliniques et environnementaux qui auraient pu aider à identifier ces facteurs. La quatrième limite est le biais introduit dans notre géolocalisation en déduisant un emplacement basé sur des emplacements fournis par l'utilisateur qui pourraient ne pas être leurs véritables emplacements. Enfin, l'inférence causale entre le revenu moyen des ménages par ville et les sujets d'intérêt des personnes vivant dans la ville géolocalisée ne peut être faite car elle est sujette à un risque de biais écologique.

À notre connaissance, il s'agissait de la première étude à recueillir des informations concernant les principales préoccupations liées au diabète sur la base des données des réseaux sociaux.

4.4 Conclusion

Nous avons démontré la faisabilité de capturer les émotions, les préoccupations et les intérêts des individus dans la vie réelle et montré que cela était un moyen efficace d'augmenter la recherche psychosociale, comportementale et épidémiologique. L'utilisation de l'analyse Twitter sur le diabète pourrait éclairer le débat public sur les problèmes liés au diabète et contribuer ainsi directement à la prise de décision publique et clinique. Les données des réseaux sociaux aideront à développer des politiques et des interventions qui prennent en compte les principales

préoccupations des personnes atteintes de diabète afin d'améliorer au final les résultats de santé publique de manière générale. Ces éléments devront être validés par des études cliniques complètes pour chaque domaine.

CHAPITRE V: DÉTECTION DE CAUSALITÉ

5.1 Méthodologie

32 millions de tweets en anglais sur le diabète collectés entre avril 2017 et janvier 2021 ont été prétraités de la même manière qu'exposée dans le chapitre IV. Les retweets et les doublons ont été supprimés. Les tweets personnels, sans blagues, ont été identifiés par un modèle de langage affiné *Bertweet,*[78] au lieu de s'appuyer sur des SVM, et finalement les tweets contenant des éléments émotionnels ont été conservés. Pour les analyses, nous avons opéré au niveau de la phrase. Dans un premier temps, les tweets contenant des informations causales (par exemple, observation, opinion, préoccupations, etc.), également appelés tweets causaux, respectivement les phrases causales, ont été détectées en ajustant un modèle *BERTweet* basé sur un ensemble de données de cause à effet étiqueté manuellement. Ensuite, plusieurs architectures de modèles ont été testées pour identifier les causes et les effets correspondants à partir des tweets causaux. Une boucle d'apprentissage active nous a permis d'augmenter efficacement les données d'entraînement. Enfin, les paires cause-effet ont été regroupées de manière semi-supervisée, décrites et visualisées d'une manière interactive.

5.2 Résultats

Les tweets causaux ont été détectés avec une *accuracy* de 71%. Le modèle le plus performant pour identifier les causes et les effets correspondants était un modèle CRF avec des features des *embedding BERTweet* atteignant une précision, un rappel et un F1 de 0,68, ce qui a entraîné 96,676 paires cause-effet détectées. Le diabète a été identifié comme le plus grand cluster agissant principalement comme cause de "décès" et de "peur". Outre le "diabète", un groupe central a été détecté dans "la mort" agissant comme un effet pour diverses causes liées au prix de l'insuline, un lien déjà détecté dans des travaux antérieurs.[79] En plus, "diabète type 1 (DT1)" et " Insuline" ont été fréquemment mentionnés. Le lecteur intéressé peut explorer le réseau cause-effet dans la visualisation interactive: https://observablehq.com/@adahne/cause-and-effect-associations-in-diabetes-related-tweets

5.3 Discussion

Dans ce travail, des modèles de langage puissants nous ont permis de détecter un grand nombre de tweets contenant des relations de cause à effet explicites et implicites ayant conduit à l'identification de 20% (96 676 / 482 583) de tweets avec des associations de cause à effet, contrairement à d'autres approches qui ont été capable d'identifier la causalité dans moins de 2% des tweets.[80] Nous pourrions éviter de définir des modèles imparfaits fabriqués manuellement pour détecter les relations causales en nous appuyant entièrement sur des algorithmes d'apprentissage automatique. Opérer sur des données de réseaux sociaux en temps réel qui s'expriment spontanément offre la possibilité d'étendre nos connaissances à partir d'une source de données alternative qui pourrait compléter les sources de données épidémiologiques traditionnelles.

Une limitation est que les relations de cause à effet sont exprimées dans les tweets et cela ne peut pas être utilisé pour une inférence causale car la source de données Twitter est incertaine et les informations partagées peuvent être des opinions ou des observations. Un autre inconvénient est que les performances de nos algorithmes causaux pour détecter les paires cause-effet ne sont pas parfaites. Cependant, le processus global et la grande quantité de données minimisent ce problème. Le manque de rappel est contrebalancé par la grande quantité de données et le manque de précision est contrebalancé par l'approche de regroupement dans laquelle les causes ou les effets non fréquents sont écartés.[81] L'amélioration de la qualité des données est certainement un point primordial à aborder pour améliorer les performances. En outre, nous tenons à souligner que les informations relatives au diabète partagées sur Twitter peuvent ne pas être représentatives de toutes les personnes atteintes de diabète.

À notre connaissance, il s'agit de la première étude ciblant à la fois les relations de cause à effet explicites et implicites sur les données Twitter liées au diabète.

5.4 Conclusion

Dans cette étude, une méthodologie innovante pour identifier d'éventuelles relations de cause à effet explicites et implicites à plusieurs mots parmi les tweets liés au diabète a été développée. La faisabilité de notre approche a été démontrée en utilisant des architectures basées sur *BERT* dans le prétraitement et la détection de phrases causales. Une combinaison de fonctionnalités *BERT* et de couche CRF a été exploitée pour extraire les causes et les effets des tweets liés au diabète, qui ont ensuite été regroupés dans une approche semi-supervisée. La visualisation du réseau de cause

à effet basée sur les données de Twitter peut approfondir notre compréhension du diabète, de manière à saisir directement les résultats rapportés par les patients d'un point de vue causal.

CHAPITRE VI: SYSTÈME D'AIDE À LA DÉCISION CLINIQUE

6.1 Méthodologie

Nous avons proposé une nouvelle méthodologie pour regrouper de manière interactive des documents biomédicaux sous forme d'arbre hiérarchique. Dans un processus itératif, un utilisateur modifie et manipule cette arborescence jusqu'à ce qu'une solution de regroupement de documents définie par l'utilisateur souhaitée soit trouvée via une interface utilisateur interactive.

Un choix de conception crucial était que les documents étaient diffusés un par un, ce qui entraîne un gain radical de consommation de mémoire. L'arborescence se compose de deux types de nœuds, les nœuds de classification et les nœuds de clustering. Un nœud de classification est un classificateur binaire d'apprentissage automatique représentant un thème ou un concept créé par l'utilisateur. Les nœuds classificateurs sont situés au sommet de l'arbre et agissent comme une barrière permettant uniquement aux documents de passer aux nœuds sous-jacents s'ils correspondent au concept défini. Le rôle d'un nœud de clustering est de scinder (cluster) des documents sur la base de *head words* identifiés automatiquement qui décrivent au mieux les documents ayant passé ce nœud.

Grâce à une interaction active semi-guidée, un utilisateur peut incorporer des connaissances du domaine au système en choisissant des données d'apprentissage appropriées pour les classificateurs via l'interface. De plus, un utilisateur explore davantage l'arbre de manière itérative et améliore ses performances de classification, en utilisant l'apprentissage actif, ce qui conduit à un regroupement plus intelligent de documents similaires et aboutit finalement à un modèle qui converge vers une solution de clustering définie par l'utilisateur souhaitée.

Le résultat de cette procédure interactive est : 1) un cadre de visualisation calibré pour un corpus de texte donné ; 2) une cascade de classificateurs dirigeant les nouveaux documents vers la branche d'arbre la plus appropriée.

6.2 Résultats

Le système a été évalué sur des résumés PubMed liés au diabète structurés de manière hiérarchique via des codes MeSH.[56,82] Le clustering hiérarchique a atteint des performances proches de l'état de l'art avec un score F1 de 73% par rapport à l'algorithme HC (HAC) dans scikit-learn[83] avec 76%. Concernant notre stratégie d'apprentissage actif, nous avons atteint une performance moyenne pondérée sur tous les codes MeSH d'un score F1 de 62% par rapport à la stratégie d'apprentissage actif de Zhang et al.[84] avec 63%. Nous avons remarqué que la consommation de mémoire pour FBE reste presque constante avec le nombre croissant de documents tandis que l'algorithme HAC croît de façon exponentielle. Cependant, le gain en consommation mémoire va de pair avec un temps d'exécution plus important.

6.3 Discussion

Un atout crucial de la méthodologie proposée est que les non-experts sans connaissances en programmation sont capables d'explorer et de cibler des sujets d'intérêt dans un corpus textuel non structuré via une interface interactive et conviviale. La transparence est profondément améliorée grâce à la visualisation des mots principaux et de la structure arborescente.[85] La transparence des systèmes d'aide à la décision clinique est essentielle pour garantir l'adoption par les cliniciens.[86] Grâce à l'interaction entre l'homme et le système, l'interprétabilité est augmentée.[87] La possibilité de creuser dans des sujets permet d'identifier des sous-thèmes et de visualiser les connaissances. De plus, l'effort d'annotation des données est minimisé grâce à la stratégie d'apprentissage actif proposée en sélectionnant les instances de données les plus percutantes. L'injection de connaissances de domaine potentielles pour conduire l'extraction de sujets est toujours une tâche difficile, que nous réalisons grâce à la participation active des utilisateurs. Par ailleurs, le caractère dynamique de notre approche nous permet d'ajouter plus de documents au fil du temps sans une reconversion complète permettant l'étude des évolutions des sujets d'intérêt.

L'interaction manuelle peut également être considérée comme une limitation, empêchant la création d'un grand nombre de classeurs. Un utilisateur a cela dit généralement une certaine idée du domaine et est capable de cibler largement les clusters souhaités ce qui évite d'avoir à définir un grand nombre de classifieurs. Nous n'avons évalué la méthodologie que sur un seul ensemble de données. Une direction future des investigations est la validation de la méthodologie sur d'autres ensembles de données pour assurer la généralisation et la portabilité dans d'autres contextes. En raison de la nature du streaming, l'arbre construit dépend fortement de l'ordre des documents, ce

qui peut affecter l'interprétation des données. Une future direction à explorer devrait être l'évaluation de l'approche sur un échantillon d'utilisateurs finaux de différents profils et niveaux d'expertise dans les techniques de clustering.

6.4 Conclusion

Une interface utilisateur interactive pour les utilisateurs sans connaissances en informatique pour l'exploration de données textuelles biomédicales non structurées en tant qu'aide à la décision clinique a été proposée. Grâce à une participation active et semi-guidée de l'utilisateur et à la visualisation des *head words*, l'algorithme converge vers une solution définie par l'utilisateur tout en améliorant la transparence. Nous avons abordé diverses limitations existantes dans les systèmes actuels de regroupement et de soutien clinique, telles que l'inclusion de la connaissance du domaine ; accroître l'interprétabilité; minimiser l'effort d'annotation manuelle par un apprentissage actif, réduisant la consommation de mémoire grâce au streaming de données et permettant ainsi la gestion de grands corpus textuels tout en atteignant des performances proches de l'état de l'art. Le système développé pourrait être bénéfique pour obtenir rapidement une vue d'ensemble sur des sujets spécifiques afin d'améliorer finalement l'exploration de la littérature dans le processus de prise de décision clinique.

CHAPITRE VII: CONCLUSION

Les contributions de cette thèse exploratoire ont été doubles: contributions épidémiologiques et techniques.

Parmi les contributions épidémiologiques, nous avons démontré la faisabilité d'exploiter les données des réseaux sociaux pour capturer les émotions et les préoccupations de la vie réelle, rapportées par les patients, ce qui représente un moyen rentable et rapide d'augmenter la recherche psychosociale et épidémiologique. En outre, nous avons démontré la possibilité d'extraire les relations de cause à effet explicites et implicites en exploitant les architectures modernes d'apprentissage automatique et le traitement du langage naturel. Cela offre la possibilité d'améliorer nos connaissances sur le diabète dans un contexte réel en exploitant les résultats rapportés par les patients.

Le débat public sur les problèmes du diabète pourrait être encouragé à l'aide des données de Twitter et ainsi avoir des implications pour les décideurs en matière de santé, les praticiens de la

promotion de la santé et la prise de décision clinique. Les données des réseaux sociaux serviront à élaborer des politiques et des interventions qui incluent les principales préoccupations des personnes atteintes de diabète, afin d'améliorer en fin de compte les résultats de santé. Pour que cela devienne réalité, nos résultats nécessitent une validation dans des études cliniques, telles que des cohortes.

En outre, nous avons abordé le manque actuel d'adoption d'outils efficaces pour analyser et extraire des informations appropriées[49–51] en proposant une interface utilisateur interactive pour les utilisateurs sans connaissances en informatique à explorer des informations textuelles cliniques non structurées pour améliorer en fin de compte la partie de résumé de la littérature dans le processus de prise de décision clinique. Un accent particulier a été mis sur la transparence, l'interprétabilité et la réduction de la consommation de mémoire. Néanmoins, davantage de temps doit être investi pour améliorer encore l'outil et le tester sur un échantillon d'utilisateurs finaux pour inclure leurs commentaires et confirmer la portabilité sur d'autres ensembles de données.

Concernant les apports techniques, cette thèse reflète bien les évolutions du traitement automatique du langage naturel. Le premier objectif utilisait des intégrations *FastText*. Avec l'avènement des plongements de mots sensibles au contexte, à savoir BERT, au cours de cette thèse, nous avons également exploité ces plongements plus puissants dans le deuxième objectif pour lutter contre la causalité. Par ailleurs, ce travail a également été un terrain de jeu pour tester différentes méthodes d'apprentissage automatique et de traitement du langage naturel et pour les valider pour l'analyse des données des réseaux sociaux. Nous avons montré que l'apprentissage actif est un moyen efficace d'augmenter les données d'entraînement et de minimiser l'effort d'annotation.

En outre, nous avons souligné que les méthodes d'apprentissage automatique peuvent être appliquées à l'épidémiologie du diabète dans l'ensemble du pipeline de flux de travail : prétraitement des données, géolocalisation des données, filtrage des données, prédiction et regroupement des informations. Et en général, cela nous a permis de cibler des ensembles de données plus volumineux avec une plus grande variabilité et donc une plus grande représentativité.

Avec le développement de notre aide à la décision clinique, nous avons abordé divers goulots d'étranglement techniques allant des limitations des méthodes actuelles de PNL, telles que la spécification préalable du nombre de sujets souhaité ou la faible évolutivité des modèles de sujets, à l'amélioration de l'effort d'annotation grâce à l'apprentissage actif pour minimiser la

consommation de mémoire. Avec cette méthodologie, nous avons contribué à l'objectif de plus d'explicabilité dans les soins de santé, en décidant consciemment de ne pas utiliser les architectures basées sur les réseaux de neurones et en proposant un système modulaire dans lequel des modèles ML plus simples sont interchangeables, pour assurer la transparence.[88]

Les futures orientations à étudier pourraient être l'extension à d'autres pays; croiser des facteurs socio-économiques plus divers avec les données des réseaux sociaux; étudier la dynamique dans les réseaux sociaux pour détecter les sujets en voie de disparition et nouvellement apparus; valider les résultats dans un cadre plus traditionnel (ex.: cohorte); explorer une méthode plus sophistiquée pour extraire les associations de cause à effet; et améliorer l'outil d'aide à la décision clinique et le valider par les utilisateurs finaux.

CHAPTER I: INTRODUCTION

Artificial intelligence transforms the healthcare sector at a tremendous speed, intervening in every corner. At the same time, with the mass adoption and integration of social media in our lives, a new data source has emerged which can be exploited for epidemiological purposes.

Diabetes mellitus is a chronic disease that affects 463 million adults (20-79 age) worldwide in 2019 and is expected to grow up to 700 million in 2045.[1]
This doctoral thesis explores artificial intelligence methods for the analysis of social media data and the development of a clinical decision support system. The use case of this thesis is diabetes and diabetes distress in the framework of the World Diabetes Distress Study.

This introduction is divided into four sections, with the first section highlighting the evolution of epidemiology from its origins to the development of modern, digital epidemiology. The second section provides an overview over diabetes epidemiology, the types of diabetes, and the diabetes sub-concept diabetes distress. A third section motivates social media as a complementary source of patient-reported data. The final section introduces the need for clinical decision support systems to support health professionals in the evidence-based clinical decision making process.

1.1 Traditional epidemiology to digital epidemiology

The roots of epidemiology date back to Hippocrates (~400 B.C.) who turned the view of disease occurrence from a supernatural to a rational one, suggesting that environmental and behavioral factors potentially affect disease development.[89] From there and until the 16th century, there was a scientific hiatus, at the bottom of which epidemiology seems to have stagnated. Indeed, its next major evolutionary step is not identifiable until 1538, when mathematical statistics was integrated through « Bills of Mortality » : weekly and yearly statistics that were produced from registers and communicated the number of deaths, as well as their causes, which one would class today under surveillance epidemiology.[90,91] In the 17th century John Graunt (1620-1674) introduced systematic

methods to examine disease occurrence and death based on these *bills of mortality* which can be considered the first modern epidemiological work.[90] These methods were improved from William Farr (1807-1883) in the 19th century to better describe epidemiological problems using statistics.[89] The 19th century represents a heyday for modern epidemiology, brought to life to a significant part by French scientists. Among them, Louis-René Villermé (1787-1863) who showed that socio-economic factors (e.g. tax, rent price, etc.) negatively correlate with mortality and thus founded social epidemiology.[90] Pierre-Charles-Alexandre Louis (1787-1872), for his part, awakened clinical epidemiology by contesting bloodletting and applying standardized methods of data collection and analysis to medicine.[92] An important pioneering work marks John Snow's investigations regarding the causes of the 19th-century cholera epidemics in London, which made him known as one of the most important contributors to modern epidemiology, and part of his work is still used today.[93] Since this moment, epidemiological methods began to flourish and many groundbreaking works have been achieved such as Doll and Hill's work linking cancer to smoking,[94] or the eradication of smallpox disease.[95] A next evolution step in epidemiology constituted the integration of molecules and genes, as risk factors for diseases, in the 1990s. Each of these evolutions were motivated by novel methodologies or data sources.[96]

The advent of the internet and its massive worldwide use, whether in social media, connected devices or e-health records, in synergy with the widespread adoption of mobile phones led to digital footprints that can be exploited for epidemiological purposes and embody the foundation for the next major transformation. The early 2000s saw a first attempt to name this new emerging area of public health combined with internet-based data: "Infodemiology".[97] Later, Salathé described this phenomenon as digital epidemiology and delivered a more sophisticated definition: epidemiology that uses data that was not generated with the primary purpose of doing epidemiology including search queries, social media data, mobile phone data or connected objects.[98] This definition implies a key strength in this novel data form, the fact that data is shared in real-time and epidemiologists are able to study patient-reported outcomes in mediums with no hierarchy between patients and doctors. Both terms "Infodemiology" and "Digital Epidemiology" coexist, but we refer in the following to the later one. An important factor accelerating the growth of digital epidemiology is the combination of two of its major ingredients: digital data and the rising strength of machine learning (ML).[98]

Traditional epidemiology suffers from vast inertia between identifying a research question, designing a study, acquiring ethical approval, including all study participants and obtaining the first research results. This process can take up to many years. Digital Epidemiology offers direct access to quickly generate research hypotheses and test those hypotheses on real-time data. Moreover, traditional epidemiology is based on data collected from healthcare providers, relying on people who have access to healthcare or decide to go to a doctor. Digital epidemiology can complement this by including people who would not go to a doctor in the first place. In addition, populations with low prevalence, such as rare disease or LGBT minorities, can be reached through the huge number of SM users.[99,100] At the same time, to leverage digital data, one depends on the access of the population to electronic devices, which are not uniformly distributed in a population creating a coverage issue.[101]

Traditional recruitment strategies, such as flyers, print or television ads, frequently create cohorts of suboptimal variability of trial participants due to self-selection bias.[102] Caplan and Friesen suggested that social media offer an opportunity to overcome existing obstacles in the recruitment process through targeted recruitment advertisements or messages.[102] As a consequence, this includes low advertising costs and reduced time to meet recruitment targets.[103–105] However, digital epidemiology may also introduce selection bias by excluding specific subgroups in each step of the selection process: if a user owns an electronic device, knows the functioning of the app, has no privacy concerns, is willing to participate, is able to use Bluetooth, and is regularly using the device.[101]

In the case of diabetes, digital technologies in interplay with artificial intelligence (AI) methods represent a major opportunity to rethink the disease as all aspects are touched, from prevention to research over care and management.[106] To further illustrate the use of digital technologies and AI for diabetes, it has been shown that mobile applications can help reduce HbA_{1c} in people with type 2 diabetes;[107] Machine-learning algorithms accurately predict personalized glycemic responses to real-life meals[108] and diagnose retinal disease;[109] Continuous/flash glucose monitoring and closed-loop systems improve glycometabolic control resulting in a lower number of hypo- and hyperglycemia.[110–112]

These digital data represent a vast potential to identify new digital markers, monitoring systems and thus, could potentially, in liaison with clinical data, improve quality of life and diabetes management, and prevent complications. Figure 1.1 visualises some of the current and future innovations for people with diabetes.

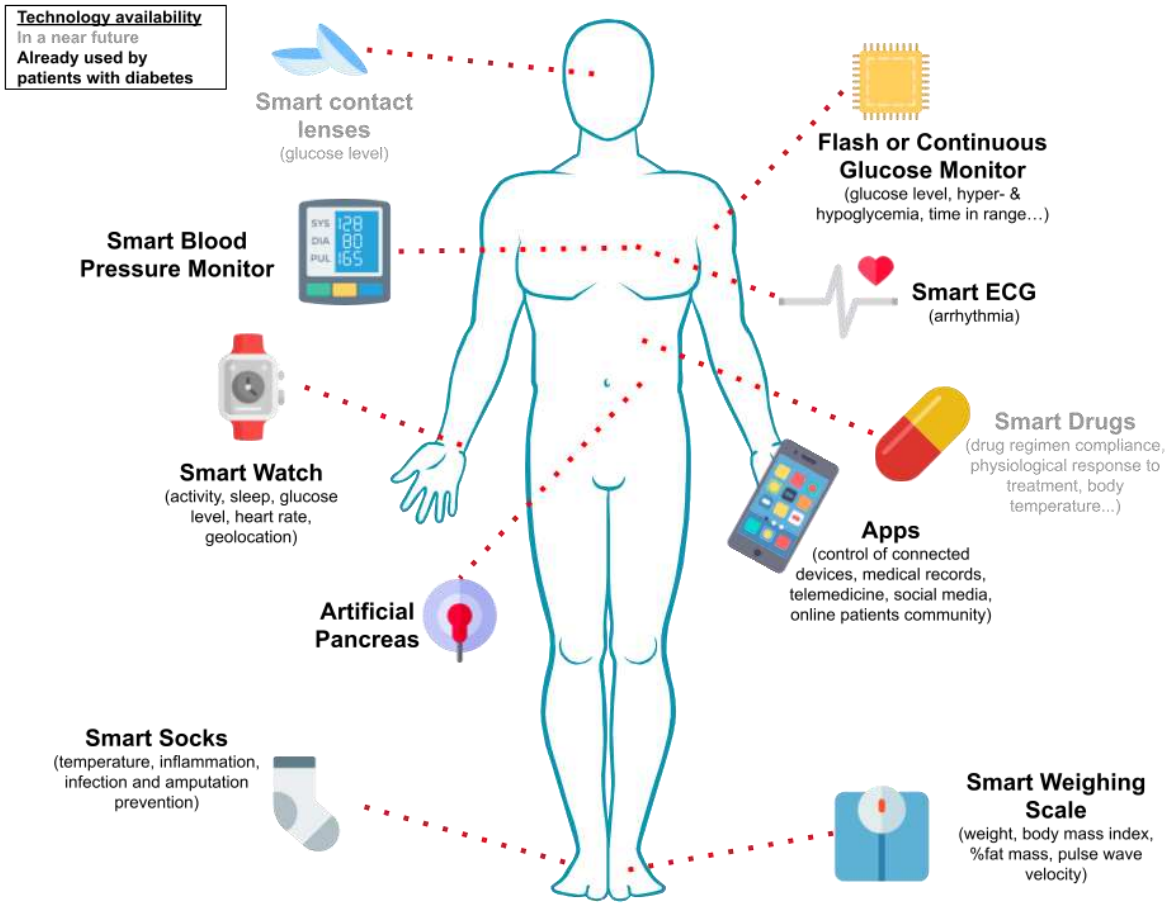


Figure 1.1: Recent and future innovations for people living with diabetes.[106]

It is crucial to emphasize the game-changing impact of artificial intelligence, not only for epidemiology but the entire healthcare system. AI is not only capable of dealing with the sheer amount of exponentially growing data but also able to detect tiny signals and extract relevant information. Various success stories have been written, from tracking disease outbreaks via mobile phones or social media,[113] the identification of skin cancer based on clinical images,[114] to electrocardiographic anomaly detection in cardiology.[115]

The combination of real-world digital data with clinical data and omics features fuelled by artificial intelligence powered systems paves the way for precision medicine to enter the era of patient-centered care.[116]

Yet, these new opportunities go hand in hand with new challenges and ethical considerations. By including digital data, clinical and research practices need to be updated to guarantee privacy by design and default, to pseudonymise and to ensure data portability which should be included from the start in a study.[116] For instance, the minimal amount of identifiable data should be used in digital epidemiology, a central principle of data protection.[117] Furthermore, transparency in research, obtaining informed consent, exchanging constantly on the different applications of the collected data and providing feedback to the community are essential factors to gain trust in these novel approaches.[116] Another risk is potential discrimination trough biases in the training data for AI algorithms which necessitates vital considerations in both the used machine learning technologies and the training data.[118] Cybersecurity plays an important role for the underlying infrastructure of digital epidemiology which is vulnerable to both cyber and physical threats.[119] Moreover, the reproducibility and replication crisis of AI technology in the medical sector poses serious issues due to unreliable design or overfitting.[120,121] To address these challenges it is crucial that AI makers, patients, health professionals and regulatory authorities work hand-in-hand to build a system that is created on *public trust* and thus benefits everyone.[118] The replication crisis could be overcome by transitioning to the open science paradigm with open sourcing data and methods with which this thesis aligns.[121]

Powerful neural network architectures require a massive amount of labeled data to detect these fine nuances and patterns in the data, a frequent bottleneck into healthcare application. To overcome this issue of scarce or expensive labeled data, the paradigm of *transfer learning* gained more and more popularity in recent years. The idea is to overcome isolated model learning and to leverage knowledge acquired for a task from auxiliary domains to make predictions for the related tasks and related domains.[122] Besides, transfer learning reduces training time significantly by relying on already pre-trained models. Thus, models have a better starting point as some knowledge is already incorporated into the system. Throughout this thesis transfer learning is exploited multiple times. For an overview over transfer learning on different data types, such as image, text, tabular, audio or time series, refer to the reviews of Cheplygina et al. and Ebbehoj et al.[123(p),124]

This thesis explores the potential of social media as complementary data source to traditional epidemiology and how complex artificial intelligence algorithms may serve and support health professionals in extracting relevant information from textual data. The focus of this work lies on diabetes research.

1.2 Diabetes

Diabetes mellitus, or short diabetes, is a chronic disease characterized by a pancreatic deficiency to produce enough insulin, a hormone regulating glucose, or if the body can not use it correctly.[125] As a result, blood glucose levels are too high, also referred to as hyperglycaemia, which in turn may lead to long-term complications, including diabetic retinopathy causing vision loss, nephropathy leading to renal failure, neuropathy provoking nerve damage with risk of foot ulcers and amputations, cardiovascular symptoms, and ultimately premature death.[125,126]

1.2.1 Insulin

Insulin is a vital hormone produced in the *beta* cells of the islets of Langerhans in the pancreas.[127] It is responsible for regulating blood glucose levels and thus plays an essential role in the development of diabetes. Insulin allows glucose in the blood to enter the cells in which it is transformed to energy. A high level of glucose in the bloodstream, for instance after food consumption, provokes the secretion of insulin in the pancreas. A potential sign for diabetes is the pancreas's failure of producing sufficient insulin or the body's inability to respond properly to it, resulting in elevated blood glucose levels (hyperglycaemia).

In the context of this thesis, mentions regarding insulin mostly refer to insulin as a drug. The proportion of patients with diabetes relying on insulin is estimated to 29.1%.[128] Primarily, people with type 1 diabetes, who do not produce insulin, or not enough, depend on external, life-saving insulin supply. Whereas it is estimated that in 2030 between 7.4% - 15.5% of people with type 2 diabetes, whose body does not properly respond to insulin, takes insulin treatment to avoid major morbidity and mortality from ketoacidosis or hyperglycaemic states and to reduce the risk of long-term microvascular complications.[129] External insulin is delivered either through insulin pens which inject the insulin via a needle or through insulin pumps which are devices providing regular insulin throughout the day.[130]

1.2.2 Type 1 diabetes

Type 1 diabetes (T1D), formerly known as insulin-dependent, is caused by an autoimmune response in which the body attacks the insulin-producing *beta* cells leading to an insufficient or non-existent insulin production.[1] The root causes of this autoimmune response are not fully known, but it is suggested that various elements, such as environmental, genetic and the immune system, interplay and thus contribute to the disease development.[2] Until today, the only existing treatment is the daily injection of insulin to maintain a stable glucose level. Type 1 diabetes affects most commonly patients in childhood, but symptoms can sometimes develop later and accounts for 5-10% of all diabetes cases.[1,3]

Frequent symptoms of the disease are blurred vision, fatigue, weight loss, excessive thirst, constant hunger and frequent urination.[1,131] Living with type 1 diabetes requires disciplined self-management including blood glucose monitoring, physical activity, healthy diet and insulin use.[2]

1.2.3 Type 2 diabetes

In type 2 diabetes (T2D) the body loses its ability to properly respond to insulin, also referred to as insulin resistance. When diagnosed, a large part of patients with T2D are asymptomatic as it develops over a long time period.[4] The origins of the insulin resistance are not entirely understood but strong risk factors are overweight, obesity, age, ethnicity and family history.[1,5]

It is the most common type of diabetes concerning around 90% of all diabetes cases worldwide and most commonly observed in older adults.[1,6] Management and prevention of T2D necessitates the identification of people with prediabetes and intervention with lifestyle changes such as weight loss, regular physical activity and healthy diet.[132,133]

Prediabetes is a state of high risk for developing diabetes with elevated glucose levels but still below the diabetes threshold.[134] It defines people with impaired glucose tolerance and/or impaired fasting glucose or increased glycated haemoglobin A1c (HbA1c) levels.[4] It signifies a risk of the future development of type 2 diabetes and diabetes-related complications.

1.2.4 Gestational diabetes

Gestational diabetes mellitus (GDM) is defined as hyperglycaemia that first develops during pregnancy and its prevalence varies greatly depending on the region (Europe: 6.1%, South-East Asia: 15%, Middle East and North Africa: 15.2%, North America: 7%).[1,7] Usually diagnosis is performed using an Oral glucose tolerance test (OGTT), in which a woman takes glucose after having fasted overnight and blood samples are taken to estimate how quickly it is cleared from the blood. Overweight, obesity, advanced maternal age, westernized diet (high intake of saturated fats and sucrose and low intake of fiber), or insulin resistance / diabetes in the family may contribute to developing gestational diabetes.[135]

1.2.5 Diabetes epidemic

The worldwide prevalence of diabetes in adults, aged 20-79, was estimated to be 463 million in 2019 (9.3% of total world population in this age group) and is expected to steadily rise up to 700 million in 2045 (10.9% of world population) according to the International Diabetes Federation, compare Figure 1.2.[1]

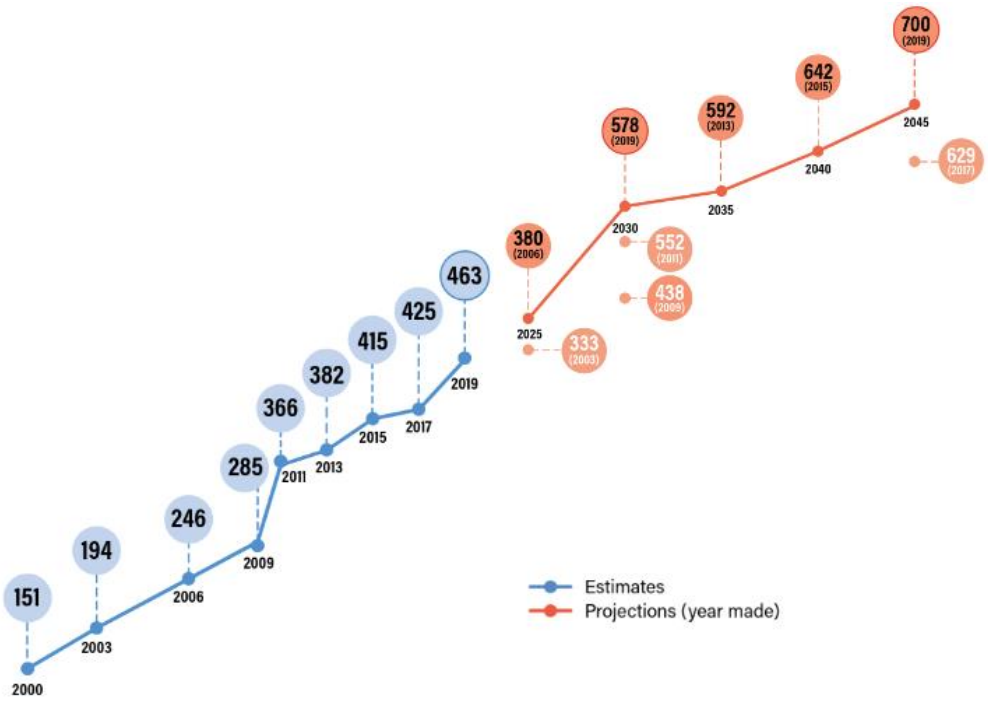


Figure 1.2: Diabetes evolution and future estimation of global prevalence in the age group 20-79 years. [1]

A large part of people with diabetes resides in low- and middle-income countries (79.4%), see Figure 1.3 for a full picture over the diabetes prevalence.[1]

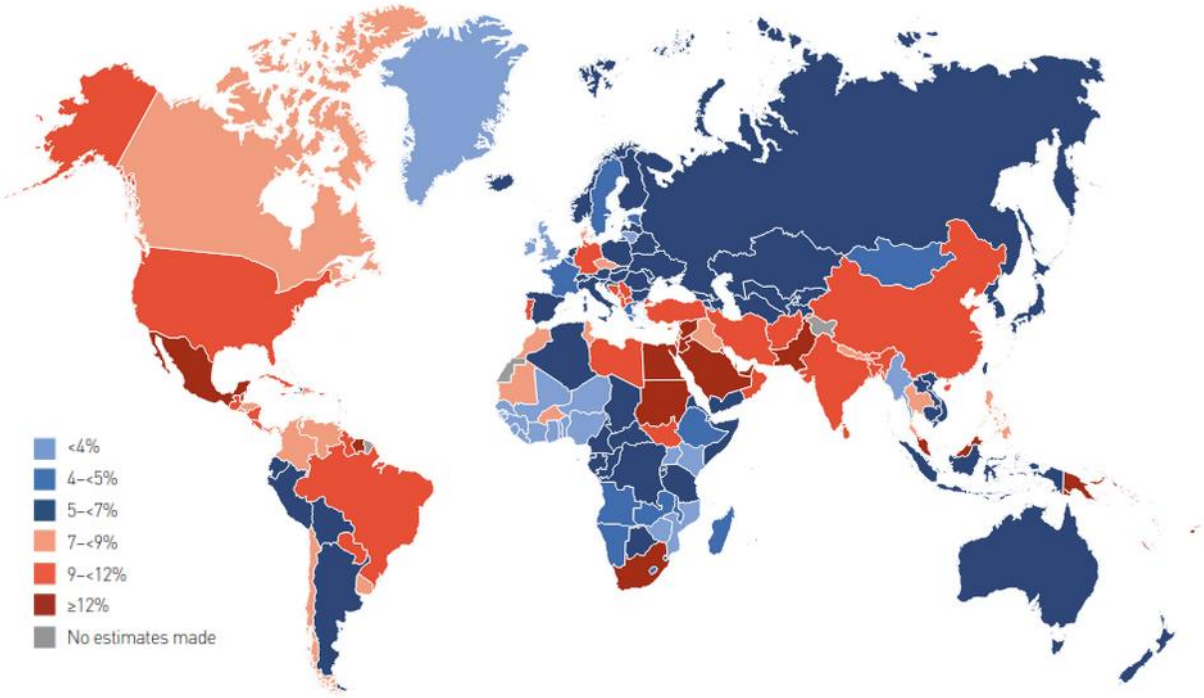


Figure 1.3: Age-adjusted comparative diabetes prevalence in adults (20-79 years) in 2019.[1]

Concerning children and adolescents (< 20 age) the prevalence for type 1 diabetes is estimated to be 1.1 million, whereas for type 2 diabetes, prevalence cannot currently be estimated.[1] In 2019 the prevalence of diabetes in men was slightly higher with 9.6% compared to 9.0% in women and generally diabetes prevalence increases with age, from 1.4% among adults aged 20-24 to 19.9% in adults aged 75-79.[1] Diabetes is responsible for roughly 4.2 million deaths resulting from diabetes and its complications.[1] In addition, in 2019 the International Diabetes Federation estimates the global health expenditures of USD 760 billion, an enormous economic burden.[1]

1.2.6 Diabetes distress

The most important psychosocial health factors in diabetes management are considered to be stress, anxiety and emotions.[136] A central sub-concept of diabetes for this thesis is diabetes distress (DD) that regroups psychological factors, such as stress, concerns, worries, emotions, fatigue and powerlessness related to the day-to-day management of diabetes.[8,9] Throughout this thesis, we also

refer to diabetes distress profiles, which encompasses all these day-to-day issues related to diabetes expressed by people with diabetes.

It is important to emphasize the distinction of diabetes distress from depression.[137] A recent scoping review from Kiriella et al. shed light on these concepts and found that despite the overlap of concepts, depression and diabetes distress are two different constructs.[138] While depression is a psychiatric condition based on symptoms and not on etiology, diabetes distress is an adaptive emotional response to disease burden with an etiology based concept (diabetes).[138] A similar characterisation describes depression as the generic feeling of depressed affect, which is not specifically connected to a condition or experience, whereas diabetes distress is positioned in the daily experience of living with diabetes.[8]

It has been shown that diabetes distress is associated with decisional conflict and for this reason affects the day-to-day disease management and the long-term risk of diabetes-related complications.[139] Moreover, it has been found that DD is directly linked to poor glycemic control and problematic self-care behaviors.[140–142] Furthermore, younger people experience greater diabetes distress.[139] Reducing diabetes-related distress may improve hemoglobin A1c and reduce the burden of disease among people with diabetes.[143]

In the literature; the two most common and validated measurements for diabetes distress are the self-reported questionnaires: Problem Areas in Diabetes (PAID) and Diabetes Distress Scale (DDS).[10,11] We still have limited knowledge about all the sources of concerns, stress and anxiety in people with diabetes including emerging ones and the current scales do not capture the full picture of DD. Limitations of these scales also include self-reporting resulting in potential inaccuracies in the data.[12]

A meta-analysis in 2017 reported an overall prevalence of 36% for diabetes distress in people with Type 2 diabetes and a significant higher prevalence in women.[13] In people with type 1 diabetes, prevalence of DD is estimated at around 20-40%.[9,14]

Diabetes is a complicated and challenging disease. It is only understandable that people with diabetes search for channels to communicate and exchange with other people in similar situations. Social media provides such a channel.

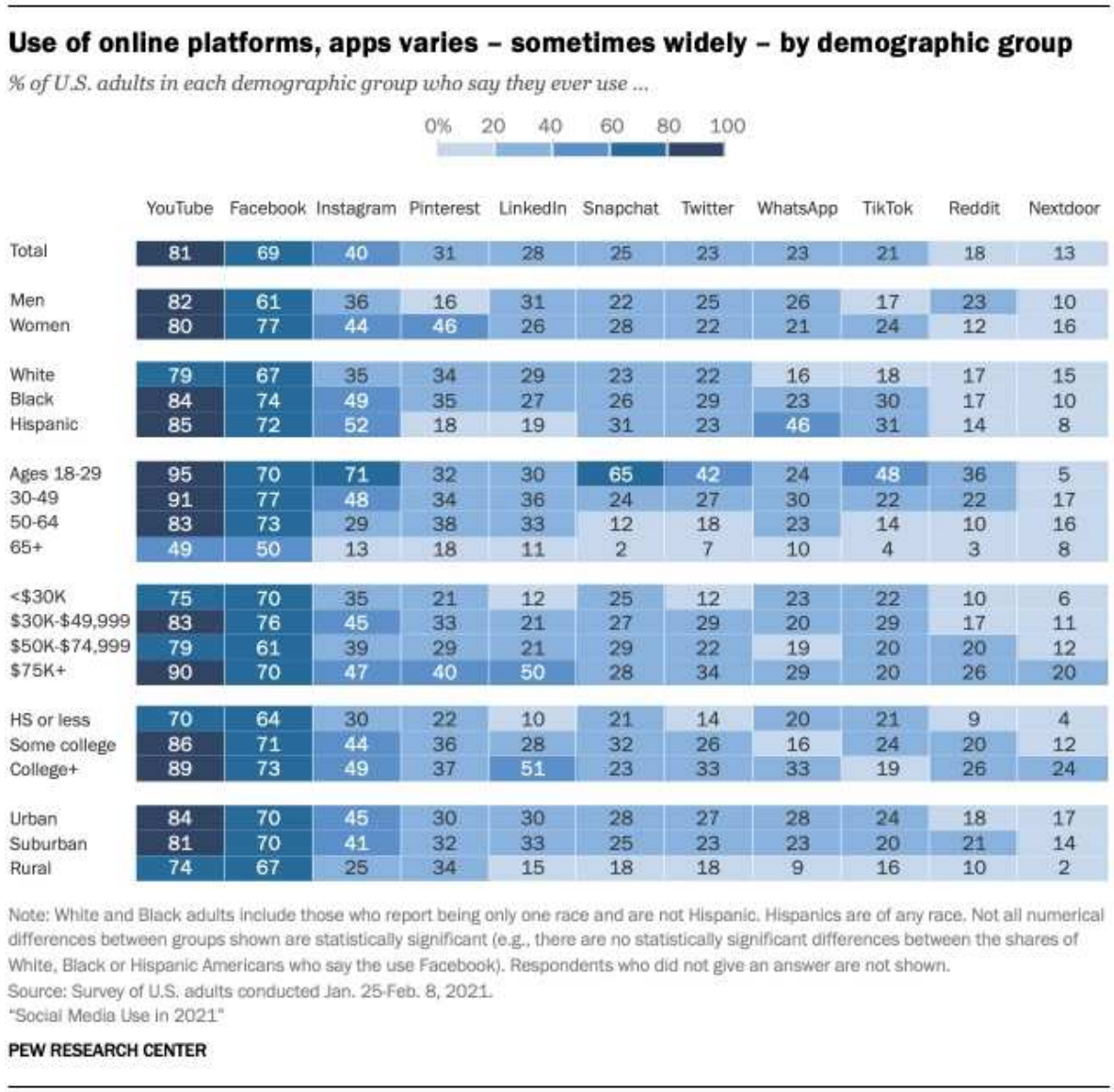
1.3 Social media

1.3.1 What is it?

The umbrella term *Social Media* (SM) emerged at the beginning of the 2000s with the growing availability of high-speed Internet access and can be defined as web-based applications whose primary function is the development and exchange of user generated content.[15,16] It conglomerates various types spanning from micro-blogging (e.g. Twitter, Tumblr), blogging (e.g. Huffington Post), social networking (e.g. Facebook, LinkedIn), online forums, media sharing (e.g. YouTube, Flickr) to Wikis (e.g. Wikipedia, Wikitravel).[144]
Social media provides a direct and easy-to-use way to exchange and network with other people compared to traditional media.

In April 2021, 4.7 billion people corresponding to 60% of the world's population were using the internet.[17] Out of those internet users, 4.3 billion were active social media users (55% of the world's population) and spending on average 2 hours 22 minutes per day on social media demonstrating the popularity and ubiquity of social media in our all lives.[17] They influence the way we live, work and communicate with each other. In a survey from 2021 among global internet users, the top 3 reasons for using social media were: "Staying in touch with friends and family" (49.7%), "Filling spare time" (36.9%) and "Reading news stories" (36.1%) underlining this omnipresence of social media in everyone's lives.[17]

In the US, the most popular online platform is Youtube with 81% users, followed by Facebook with 69% and Instagram (40%), compare Figure 1.4.[19] Twitter lays in the midfield with 23% active users and higher popularity among the 18-29 years compared to older people.[19] Younger people prefer Snapchat, Instagram, Facebook and Youtube, whereas older people generally are less present in other social media than Facebook and Youtube.[19] ANNEX 1 displays a statistic of the share of internet users visiting social network sites in selected countries.



Use of online platforms, apps varies – sometimes widely – by demographic group

% of U.S. adults in each demographic group who say they ever use ...

0% 20 40 60 80 100

	YouTube	Facebook	Instagram	Pinterest	LinkedIn	Snapchat	Twitter	WhatsApp	TikTok	Reddit	Nextdoor
Total	81	69	40	31	28	25	23	23	21	18	13
Men	82	61	36	16	31	22	25	26	17	23	10
Women	80	77	44	46	26	28	22	21	24	12	16
White	79	67	35	34	29	23	22	16	18	17	15
Black	84	74	49	35	27	26	29	23	30	17	10
Hispanic	85	72	52	18	19	31	23	46	31	14	8
Ages 18-29	95	70	71	32	30	65	42	24	48	36	5
30-49	91	77	48	34	36	24	27	30	22	22	17
50-64	83	73	29	38	33	12	18	23	14	10	16
65+	49	50	13	18	11	2	7	10	4	3	8
<$30K	75	70	35	21	12	25	12	23	22	10	6
$30K-$49,999	83	76	45	33	21	27	29	20	29	17	11
$50K-$74,999	79	61	39	29	21	29	22	19	20	20	12
$75K+	90	70	47	40	50	28	34	29	20	26	20
HS or less	70	64	30	22	10	21	14	20	21	9	4
Some college	86	71	44	36	28	32	26	16	24	20	12
College+	89	73	49	37	51	23	33	33	19	26	24
Urban	84	70	45	30	30	28	27	28	24	18	17
Suburban	81	70	41	32	33	25	23	23	20	21	14
Rural	74	67	25	34	15	18	18	9	16	10	2

Note: White and Black adults include those who report being only one race and are not Hispanic. Hispanics are of any race. Not all numerical differences between groups shown are statistically significant (e.g., there are no statistically significant differences between the shares of White, Black or Hispanic Americans who say the use Facebook). Respondents who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.
"Social Media Use in 2021"

PEW RESEARCH CENTER

Figure 1.4: United States Demographics over different social media.[19]

1.3.2 The case of Twitter

A social media platform this thesis is centered around is Twitter, a micro-blogging platform with 396 million active users in April 2021 and 500 million tweets sent each day.[17,20]

From all micro-blogging platforms, Twitter is by far not the most popular one given the number of active users. As a matter of fact, Twitter lists only at rank 16 of the social media platforms with most active users with Facebook, Youtube and WhatsApp at the top globally, as compared in Fig 1.5.[17]

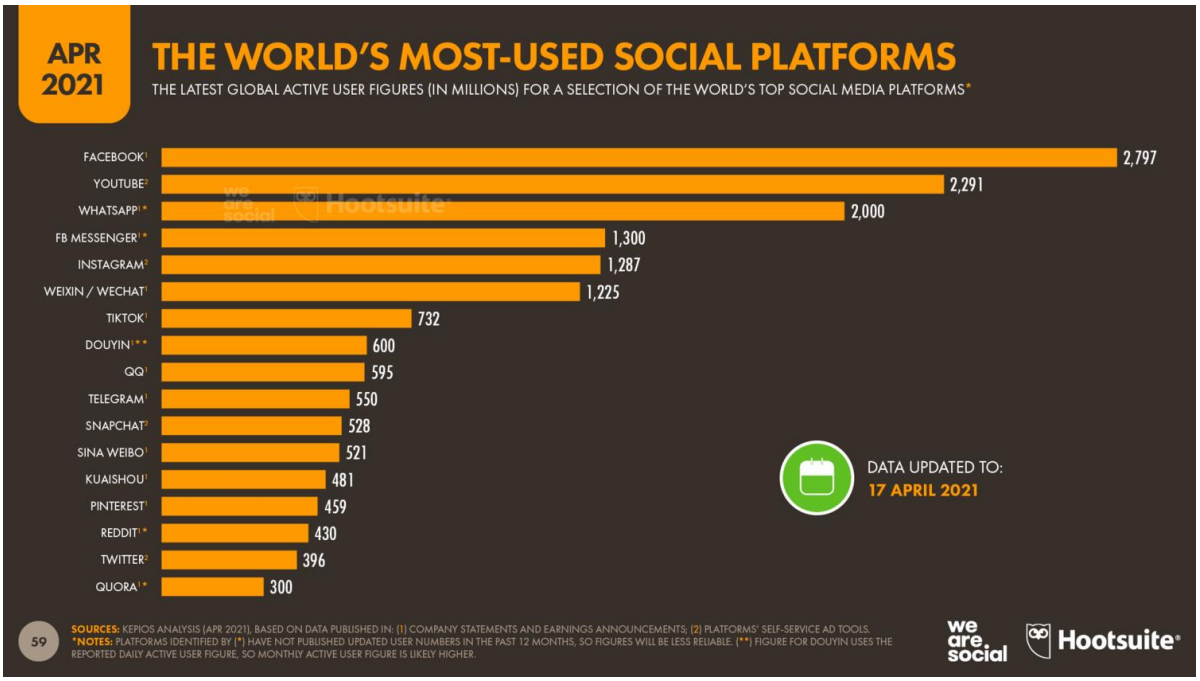


Figure 1.5 Worldwide active users over different social media platforms.[17]

A profound distinction of Twitter to its other social media counterparts is its public character. Commonly, the information shared on Twitter is public and potentially searchable by everyone worldwide, including people without a Twitter account.[21]

Simple interactions shape the Twitter world such as following each other (i.e. receiving more content from the respective person), replying, liking and retweeting (i.e. sharing the content of another person). These features allow researchers to create graphs to conduct network analyses.

Globally, the majority of Twitter users is male (63.7%) and ages under 50 with 17.1% of users being 18-24 years old, 38.5% being 25-34 and 20.7% being 35-49 years old.[17] A survey on U.S. adult Twitter users from 2019 showed their high education level with 42% having graduated from college contrary to 31% in the general U.S. population (Figure 1.6).[22] Besides, Twitter users are over-proportionally often situated in the high income class with 41% earning more than $75,000 in comparison with 32% of the general population.[22]

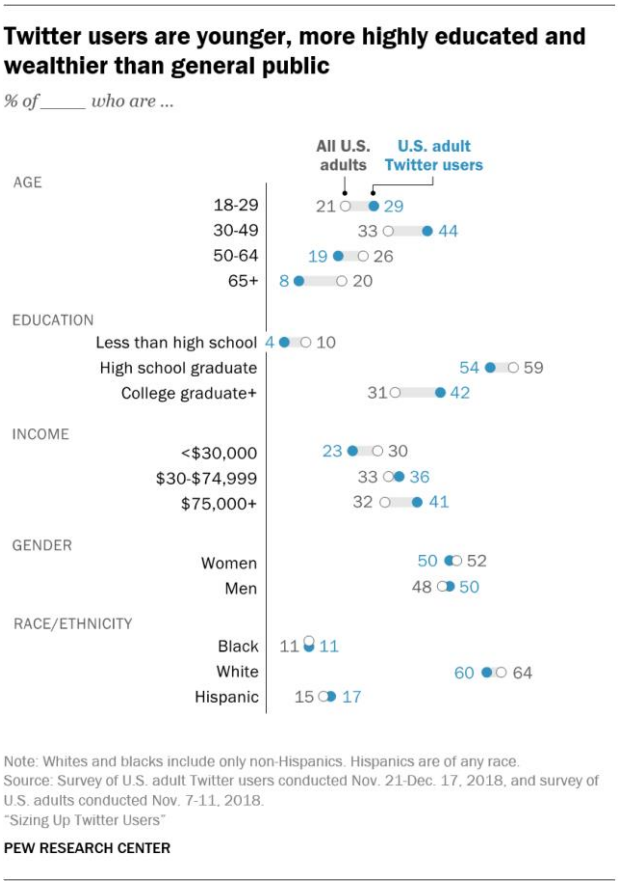


Figure 1.6: Socio-economic characteristics of Twitter user.[22]

Figure 1.7 provides an overview over the leading countries in terms of the number of Twitter users. Unsurprisingly, the USA dominates with 73 million users, followed by Japan with more than 55 million and India with 22 million users.

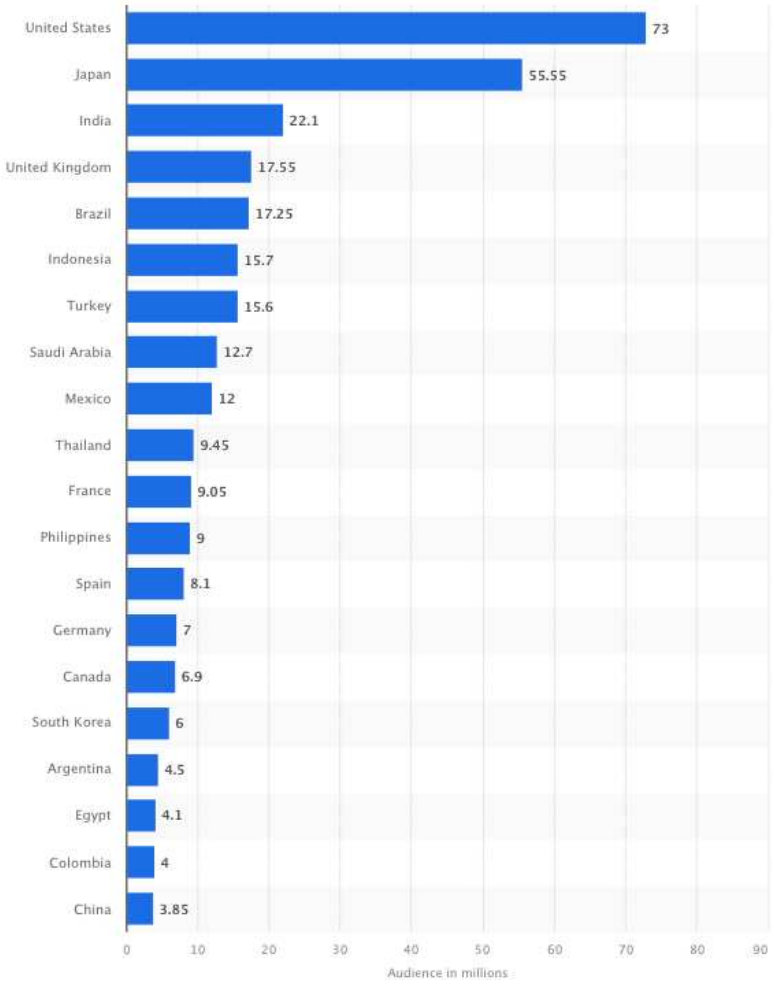


Figure 1.7: Leading countries based on number of Twitter users as of July 2021 (in millions).[145]

When it comes to analyzing Twitter data, a high importance lies on the preprocessing of the tweets. Usually, tweets are noisy, unstructured, containing slang or misspelled words. A key with successfully extracting relevant information lies in the right preprocessing and analyses techniques. Natural language processing (NLP) methods are predestined to fulfil this task and gap the bridge between social media and public health research.

1.3.3 Social media for public health research

In recent years the number of researchers, studying Twitter data for health purposes, has been growing tremendously. The public character of Twitter and in consequence easy data access is certainly the main reason for its attractiveness to researchers and its dominance over other social media.[23] While specific health related forums and blogs exist, such as TuDiabetes or Beyond Type 1

in the case of diabetes,[146] they appear to be less popular for health research compared to Twitter or Facebook.[28]

Sharing and searching health information in social media is already a common practice.[147] A survey from the Pew Research Center confirmed that social network users exchange about personal health experiences with friends or in groups, raise money to draw attention to a specific health issue or were looking for health information.[24]

Social media is continuously evolving. Where a principal attribute of social media in relation to other applications is the possibility to interact with other users via user-generated content (e.g. videos), it nowadays includes aspects of patient empowerment or participatory medicine.[148] Thus, social media triggers the transformation of the healthcare system towards a patient-centered model enabling patients to improve their self-management and socialize with each other.[148] Tracking those health-related information, treatment and feelings related to posts or discussions on social media offer new opportunities to develop methods to improve healthcare.[25,27,149]

Bour et al. provided a detailed review over social media use for health research.[28] They observed the evolving character of social media use in health research from concentrating on communicable diseases (e.g. influenza, HIV)[29,30] in earlier studies to most recently including recruitment strategies[150] and data collection for infoveillance (syndromic surveillance using the internet).[28,31]

Various health-related actors benefit from social media in one way or the other. Health professionals and researchers stay informed about the latest scientific publications, remotely follow medical conferences, interact with patients for education purposes, or even use SM as a tool to recruit study participants.[105,151–153] Patients are able to join virtual communities to connect with peers suffering from similar conditions, to exchange about health related information and to provide emotional support.[154,155] It also enables patients to be more engaged in their care and take a central role.[156] Physicians are able to improve their knowledge and abilities with social media.[157] Thus, social media is a digital space with flat hierarchies between the participants enabling the access and sharing of medical information about a patient's health and especially their feelings and emotions, which would not have been possible to collect in a face-to-face setting with a healthcare professional.

However social media data should be used with caution. Poor quality and reliability of social media represents hurdles in its usability.[32] Contrary to evidence based medicine which de-emphasizes anecdotal stories, social media tends to emphasize them.[33] Another major issue is the question of

privacy. Patient privacy breaches can potentially cause much more harm on social media compared to face-to-face with the healthcare professional due to the open character of SM and the permanency of digital information.[34] To preserve privacy concerns, patients who desire to communicate with their healthcare provider should be made aware of those potential privacy breaches.[158] Chretien and Kind also warn of a decline in empathy that might come with increased time spent interacting online with physicians and so losing the ability to relate to each other.[158] Further ethical issues arising with SM for health purposes are getting consent of online users, in particular for adolescents or their parents, and preserving the anonymity of study participants.[35] Denecke et al. further point out that digitally tracking personal behaviors in patients, such as checking if patients really maintain a healthy diet, may threaten the trust in the patient-physician relationship and as a result influence the patient treatment.[35] For the moment, no official general guidelines exist to deal with those ethical issues on social media.[35]

1.3.4 Social media for diabetes research

Social media, such as Twitter, are increasingly popular among people with diabetes.[36] In a qualitative analysis of various SM platforms, researchers showed that people with diabetes share practical tips, connect in groups with each other and share humor and pride.[37] However this was carried out on a small dataset of roughly 500 social media pieces. Nevertheless, asking people with diabetes directly confirmed that the diabetes online community (DOC) represents a strong pillar of support, where people encourage each other and strong engagement in online communities correlates with better glycemic levels.[146] Another review emphasizes the beneficial role of diabetes online communities with relatively few negative consequences.[38] Potentially, these platforms provide a form of social support for people with diabetes which seems to be a main driver in therapy adherence.[39] A specific example for the positive effect of the diabetes online community is the movement #WeAreNotWaiting which raised awareness of deficits in glucose monitoring.[159]

Beguerisse-Diaz et al. studied topics that arise in 2.5 million tweets over a period of 8 months, using co-occurrence graphs and the topic model *Latent Dirichlet Allocation (LDA),* and found thematic groups such as health information, news, social interaction, commercial, humorous tweets.[40,160] Using network analysis, they identified the most influential users to be bloggers, advocacy groups and NGOs related to diabetes.[40]

Interventions on social media, through technological devices (e.g. mobile phones) for health education and awareness programs, have a beneficial impact on reducing HbA1c and raise high satisfaction in patients.[41,42] Petrovski and Zivkovic support the positive effect of social media use on lowering HbA1c and recommend it as an additional communication tool for adolescents and young people with type 1 diabetes.[43] A survey showed that visiting social media sites correlates with an improved self-management behavior.[44]

The effect of social media interventions on health-related quality of life is controversial with some studies reporting an improvement,[161,162] while others did not find any difference regarding quality of life between the groups after the intervention.[163,164]

Crowdsourcing to gather information from the diabetes online community is a useful application as well. For instance, social media has been used for diabetes device safety surveillance, where participants reported information about blood glucose monitors, continuous glucose monitors, insulin delivery devices, insulin pumps and insulin pens, or for monitoring health risk of patients such as hypoglycemia.[45,46]

Concerning diabetes distress there is still a lot we do not know.[8] An important point in this thesis is the investigation and identification of diabetes distress patterns on social media.
Frequent limitations in diabetes-related social media studies are the limited sample size, and self-reported behaviors.[44]

Social media is one major type of textual data, but more generally textual data in the health care setting are embedded in electronic health records (EHR), free text in questionnaires, doctor-patient interactions or scientific literature. With the exponentially growing biomedical text, intelligent methods are required to extract relevant clinical information from this free-text data.[165]

1.4 Natural language in clinical decision support

This rapidly expanding mountain of text data makes it challenging for health professionals to analyze and extract valuable information. In the present health system, health professionals frequently face situations in which they often do not know that specific patient data are accessible

somewhere, do not know how to access it, do not have time to look for it or are not up-to-date with the latest state-of-the-art in medical research.[47,48] An extreme example illustrating the problem of exponentially growing literature is the case of COVID-19. In May 2020 the number of published articles in the PubMed database with the term "COVID-19" in the title or abstract indicated 7440, whereas only one month later this number rose up to 17.559 articles resulting in on average 137 new publications per day.[166] Indeed, independently of COVID-19, the number of publishing authors and publications is both increasing by around 3% a year according to a report issued by the International Association of Scientific, Technical and Medical Publishers.[167]
This sheer amount of data makes it almost impossible to be read and analyzed by single health professionals and in consequence to exercise evidence-based medicine (EBM) properly.[165,168] EBM combines clinical experience with the value of the patient and the best available research information to guide clinical decision making.[169]

Intelligent clinical decision support systems (CDSS) provide a possible way out of this dilemma. They provide the capability to efficiently filter those data, extract the relevant information and present them in an understandable manner. Even beyond information extraction from literature they provide alerts, prescribe recommendations, image interpretation and diagnostic assistance.[51] However, bottlenecks hampering wide adoption of these tools in clinical practice are poor user-interfaces, the failure to integrate those tools naturally in the health care professional's routine; the design of those tools from a too strong technical perspective or a lack of physician acceptance.[49–51] In a report, Bates et al. shared 10 learnt lessons in implementing clinical decision support systems in which they point out the importance of user-friendliness of the tool, making it easy for health professionals to "do the right thing".[52] Furthermore they emphasized that managing and maintaining the systems is critical for the successful delivery of decision support.[52]

Another important reason for the limited adoption, is the nature of machine learning algorithms integrated in those systems, often referred to as "black box models", and their lack of interpretability for both physicians and engineers who developed them.[53,54] In contrast, explainable artificial intelligence (AI) includes interpretable AI models where strengths and weaknesses of a decision-making process are transparent.[170] Wasylewicz and Scheepers-Hoeks also see a great and promising challenge to combine healthcare with big data, but warn from "Black box" systems and emphasize the need for validation and acceptance of those systems.[47] Simon et al. found in 2008 that only less than 1 out of 5 practices implemented some sort of decision support.[171] As an answer

to this little adoption, a roadmap on clinical decision support was published from the American Medical Informatics Association (AMIA) identifying three principal directions to tackle to ensure broad use of CDSS: 1) Best knowledge available when needed; 2) High adoption and effective use; 3) Continuous Improvement of Knowledge and clinical decision support systems methods.[55]

In this thesis we aim to examine the first direction of how to best structure, analyze and present knowledge and how known gaps hampering adoption of clinical decision support systems can be addressed such as interpretability, ease-of-use and handling large data corpora.

CHAPTER II: OBJECTIVES

Addressing a research question in traditional epidemiology is mainly hypothesis driven. A hypothetical idea of a relation between entities is postulated and confirmed or denied by exploiting data related to the research question. Contrary to this strategy, in this work, a data-driven approach is followed. A priori no hypothesis is stated on the data used in the different objectives. However, the fundamental hypothesis of this work is that social media data, such as Twitter, is a useful complementary source of information compared to traditional health research data to better capture what are the main issues of people with diabetes.

The contributions in this work are two-fold. On the epidemiological side, this exploratory thesis aims to extract information out of diabetes-related data which could then be used in a next step to formulate hypotheses which are studied in a more traditional setting (e.g. cohort study). On the technical side, the aim was the development and validation of innovative machine learning algorithms, which were open-sourced to encourage the community to reuse them for their research work.

2.1 Diabetes distress profiles

Following this data-driven approach, the first objective of this thesis was to identify diabetes distress profiles and diabetes-related concerns using social media data. To the best of our knowledge this was the first study capturing information about key diabetes-related concerns using social media data. Chapter IV will outline this objective.

2.2 Causal relationships

Currently, it is challenging to automatically identify *cause* and corresponding *effect* relationships in textual data. In the second objective of this thesis, the detection of *cause-effect* relationships in diabetes-related tweets was studied leveraging powerful machine learning algorithms. Similarly to

the first objective, this is the first study addressing the identification of *cause-effect* pairs in diabetes related tweets. Chapter V will detail the second objective.

2.3 *Feedbackexplorer* - Clinical decision support

Health professionals practice evidence-based medicine on the basis of the best available knowledge. The exponentially growing amount of biomedical data (literature, EHR, social media) poses serious difficulties to properly exercise EBM based on recent, up-to-date information. The third objective tackled the literature summarization part in the clinical decision making process and dealt with the development of a methodology to efficiently structure, analyze and visualise scientific literature with a focus on interpretability, ease-of-use for health practitioners without programming skills and the capability of handling large data corpora. Chapter VI will illustrate the last objective.

CHAPTER III: GENERAL CONCEPTS

In this chapter, the main data sources and principal methodological concept, transversal to thesis, will be outlined. This will provide a global intuition over the concepts of natural language processing and machine learning methods used and how these methods were evaluated. The following chapters of this thesis, detailing the studies conducted, will each have a separate section dedicated to the specific methodology applied to the study and refer, for the main concepts, to this general introduction.

The first section focuses on the data used in this work, namely Twitter data for the first two objectives addressing the identification of diabetes distress related patterns and causality in tweets and PubMed abstracts to evaluate the third objective.
The following subsection addresses a wide variety of concepts applied starting with an introduction on machine learning methods, explaining the data representation and in particular the principle of word embeddings, detailing supervised and unsupervised algorithms used, their evaluation and finishing with introducing active learning.

3.1 Definitions

In the following some initial definitions are provided based on the reference book of Baeza-Yates and Ribeiro-Neto:[172]

- **Token/Word**: a string (sequence of characters)
- **Document:** single unit of information, typically text. In this context, a *document* can be a Tweet or PubMed abstract
- **Index term**: keyword (or group of related words) which has some meaning of its own. In chapter VI we refer to this also as *head words*
- **N-gram:** Contiguous sequence of N items from a text sample
    - Example: 3-grams of “Apple” are: <”App”>, <”ppl”>, <“ple”>
    - Example: The 3-grams of “James Dean is a great actor” are:

< “James Dean is”>, <”Dean is a”>, <”is a great”>, <”a great actor”>

- **Phrase/Sequence:** Sequence of ordered tokens
- **Stopwords**: words which are very frequent and do not carry meaning (e.g. “the”, “a”, “of”)
- **Tokenization:** Process of separating a chunk of continuous text into separate words
- **Lemmatization:** morphological analysis of words trying to map a verb form to infinite tense and nouns to a single form so that they can be analyzed as a single item. For instance, the lemmatization of “walked” is “walk”.
- **Corpus/collection:** collection of documents
- **(Word) Embedding:** Encoding of a word or document into a fixed-size (high dimensional) vector representation
- **Vocabulary:** Number of distinct words in a document/collection

3.2 Data

3.2.1 Twitter data

The real-time micro-blogging social network Twitter is the primary data source for the first and second thesis objective.

3.2.1.1 Twitter’s API

Access to the data is granted via Twitter’s application programming interfaces (APIs) endpoints, which provide broad access to the data by streaming public Tweets from the platform in real-time. Specifically for this work, the *filter stream* endpoint was accessed enabling users to define keyword-based filter rules and return the matching tweets from a random sample of 1% of all tweets. As our work was centered around diabetes, we defined keywords like *diabetes, insulin* or *hypoglycemia.* For a full list of keywords, please see Table 3.1. In the framework of the World Diabetes Distress Study keywords in various other languages were defined to filter tweets from different countries.

The Twitter APIs returns Tweet objects encoded in the JavaScript Object Notation (JSON) format, which stores the tweet's information based on key-value pairs with named attributes and associated values.

All data collected in this work was publicly posted on Twitter. According to the privacy policy of Twitter, users agree to have this information available to the general public.[21]

English keywords		
glucose	insulin	blood glucose
#glucose	#insulin	#bloodglucose
insulin pump	diabetes	t1d
#insulinpump	#diabetes	#t1d
type 1	t2d	type 2
#type1	#t2d	#type2
#bloodsugar	#dsma	#type2diabetes
#doc	#bgnow	#wearenotwaiting
#insulin4all	dblog	diyps
hba1c	#dblog	#diyps
#hba1c	#cgm	#freestylelibre
diabetic	#gbdoc	freestyle libre
#diabetic	#gdm	finger prick
gestational diabetes	continuous glucose monitoring	#fingerprick
#gestational	#continuousglucosemonitoring	#changingdiabetes
#thisisdiabetes	#lifewithdiabetes	#diabetesadvocate
#stopdiabetes	#diabadass	#diabetesawareness
#diabeticproblems	#justdiabeticthings	#t1dlookslikeme
#diaversary	#diabetestest	#t2dlookslikeme
pwd	#duckfiabetes	#GBDoc
#pwd	#kissmyassdiabetes	

Table 3.1: English keywords used to collect tweets from the Twitter API

3.2.1.2 Tweet description

Principally, a tweet is a short text message posted on the social network Twitter with a maximum number of 280 characters. A sample tweet is shown in Figure 3.1. The default is to post a tweet publicly, but it is also possible to switch to private mode which makes the tweet only visible to one's followers. A tweet is accompanied by a large amount of metadata, including images, videos, user information or geolocation information. The most important meta-data fields for this thesis are summarized and explained in Table 3.2.

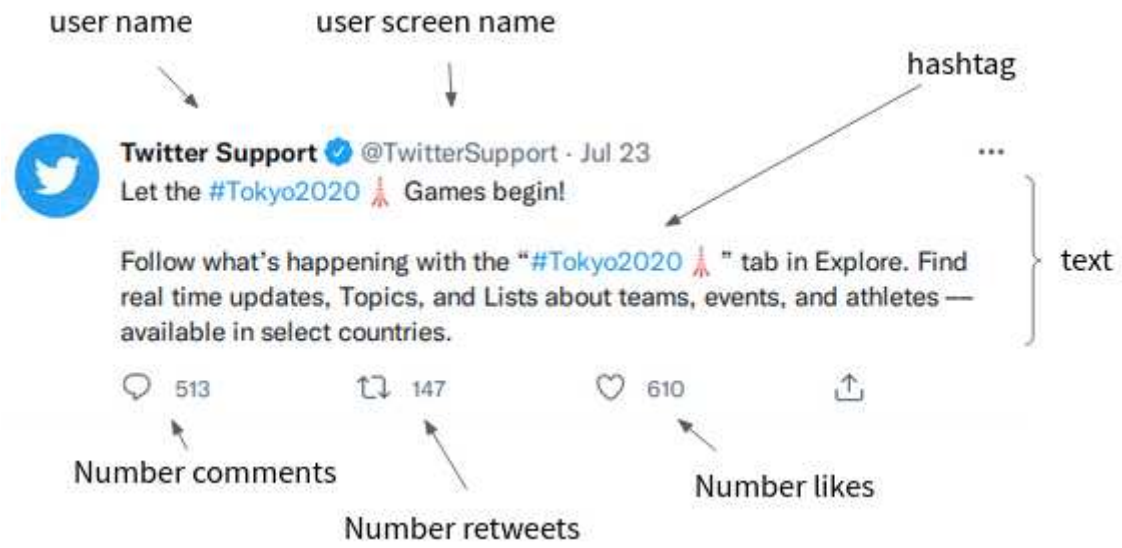


Figure 3.1: Sample Tweet

Attribute	Attribute child	Description	Example
created_at		UTC time of Tweet creation	Fri May 12 11:34:23 2019
id		tweet's unique id	149385738438
text		Tweet text	*hello, I am an example*
user	id	user id	*1234567*
	name	user name	*Twitter Name*
	screen_name	screen name and handle	*twittername*
	location	user-defined location for this account's profile	*Los Angeles, CA*
	description	user-defined description about the account	*I am a teacher living in Los Angeles*
place	full_name	Full representation of the place's name with Twitter's geo-tagging	*San Francisco, CA*
	name	Short representation of the place's name with Twitter's geo-tagging	*San Francisco*
	country_code	Shortened country code	*US*
entities	hashtags	Represents hashtags in the Tweet text	*#newFeatures*
	media	Represents media elements uploaded with the Tweet	*images*
	urls	Represents URLs included in the text	*https://developer.twitter.com*
lang		language identifier of the Tweet text	*en*

Table 3.2: Meta-data fields most relevant for this thesis. Examples are fictive

3.2.2 Pubmed diabetes-related abstracts

Abstracts were downloaded from PubMed, a large biomedical bibliographic database with more than 33M references[1] to biomedical and life sciences journal articles starting from 1946 and provided by the National Library of Medicine.[56] A specific characteristic of PubMed abstracts is their categorisation via MeSH (Medical Subject Headings) codes. The MeSH codes are structured hierarchically and used for indexing, cataloging and searching biomedical documents. Each MeSH code is composed of the MeSH descriptor name, a number indicating the location in the hierarchical tree and a unique identifier. Figure 3.2 shows the structure of diabetes-related MeSH codes, the MeSH codes relevant for this thesis. For instance, the MeSH code "Diabetes Mellitus" has the tree

[1] latest statistics: https://pubmed.ncbi.nlm.nih.gov (accessed: 29-09-2021)

location “C19.246” and unique identifier “D003920”. One of its children “Diabetes Mellitus, Type 2” extends the tree location to “C19.246.300” and has the unique identifier “D003924”.

The third objective of this thesis will use Pubmed abstracts of publications related to diabetes and in particular takes advantage of its hierarchical structure to evaluate the developed methodology.

Figure 3.2: Diabetes related MeSH structure.[2]

3.3 Machine learning and natural language processing

Machine learning is a branch of artificial intelligence designed to learn from the surrounding environment and experience.[57] These algorithms “learn” to achieve a desired task through repetition based on input data which is referred to as the training data. Once the training of an algorithm is finished, it is capable of generalizing and processing unseen data with respect to the desired task.

[2] Tree view taken from: https://meshb-prev.nlm.nih.gov/treeView

A specific architecture of machine learning algorithms are Neural networks (NN) or Artificial Neural Networks (ANN) which have been inspired by efforts to simulate the functioning of the brain with neurons, being the nodes, and connections between them.[58] Each neuron receives some inputs, usually multiplied by a weight factor, which are then aggregated and passed through an activation function to produce an output. In *multi-layer* or *feed-forward networks,* neurons are arranged in a layered manner in which several hidden layers are placed between input and output layer. A neural network architecture having multiple hidden layers stacked upon each-other is commonly referred to as deep learning.[59] For a broad introduction of neural networks, please refer to the report of Schmidhuber.[173]

Human language is complex, filled with ambiguities and subject to a person's interpretation making it tremendously difficult to develop programs which reliably determine the intended meaning of the text data. Natural language processing (NLP) explores how computers can be used to understand and manipulate natural language.[60] Naturally it lies in the intersection of computational linguistics - rule-based human language modeling -, computer science and mathematics. These approaches jointly enable computers to process human language and to *understand* its full meaning including a person's sentiment and intent. NLP addresses various tasks such as translation, summarization, entity recognition, sentiment analysis or information extraction.
A key task in natural language processing is language modeling which aims to determine the probability of a given word or span of words in a sentence using various probabilistic and statistical techniques. The first neural language model was introduced by Bengio et al. as a simple feed-forward neural network which learns the parameters of the conditional probability of the next word given all the previous ones.[174]

Throughout this thesis, the machine learning, neural network and NLP concepts that will be applied are introduced in the following sections.

3.4 Data representation

Machine learning algorithms usually work with numerical data and prefer well defined fixed-length inputs and outputs, whereas textual data is generally messy. Thus, textual data, such as social media data, electronic health records or clinical notes need to be transformed to numbers.

An early model capturing the relative importance of the terms in a document is the *vector space model* introduced by Salton et al.[175] The basic idea is to represent each document in a collection as a point (vector) in a common vector space with the property that points close to each other in this space are semantically similar.[176] The values of the elements in a point are derived from frequencies, such as the number of times a given word appears in a given context.[176] *Vector space models* form the basis for later developments in natural language processing.

Two key breakthroughs drove the boom in natural language processing in recent years and revolutionised the way textual data is reshaped into useful numbers: first, the ability to generate meaningful fixed-size vector representations of words, so called *word embeddings* or *dense vectors*, triggered by the method *word2vec*[61,62] and extended by *Glove*[177] and *FastText;*[63] and secondly, to enhance beforehand mentioned methods by the ability to generate context-aware word and sentence representations[64] using a *Transformer* architecture.[65] The *Transformer architecture* is the key innovation in the popular *BERT* model.[64]

Word embeddings build the foundation of all three objectives and the *Bert* model plays a crucial role in our second objective to identify causal associations. For this reason, these methods will be introduced in the following sections. In addition, readers, who are less familiar with the NLP and ML, shall gain an intuition about *word embeddings* and the way we deal with text data.

3.4.1 Word embeddings

Simply spoken, a *word embedding* is a fixed-size, dense vector of real numbers. Usually, they capture context and word relations in a way that words that are similar semantically for humans also have similar word embeddings and thus are close in their vector space.

There exist several different techniques for constructing word embeddings. The aforementioned first revolution in NLP started in 2013 with the work of Mikolov et al.[62] who introduced the *Word2Vec* method, based on a shallow neural network, composed of an input layer, one hidden layer and one output layer. They calculated word embeddings on a large amount of unlabeled text data and revealed interesting semantic and syntactic relationships. For instance, it was shown that the vector representation of the word *Queen* can be derived by the simple algebraic vector operation: vector(*King*) - vector(*Man*) + vector(*Woman*) and can be interpreted as the analogy "king is to queen as man is to woman".[178] Figure 3.3 illustrates further examples of word embeddings in the vector space and the idea of semantic similarity.

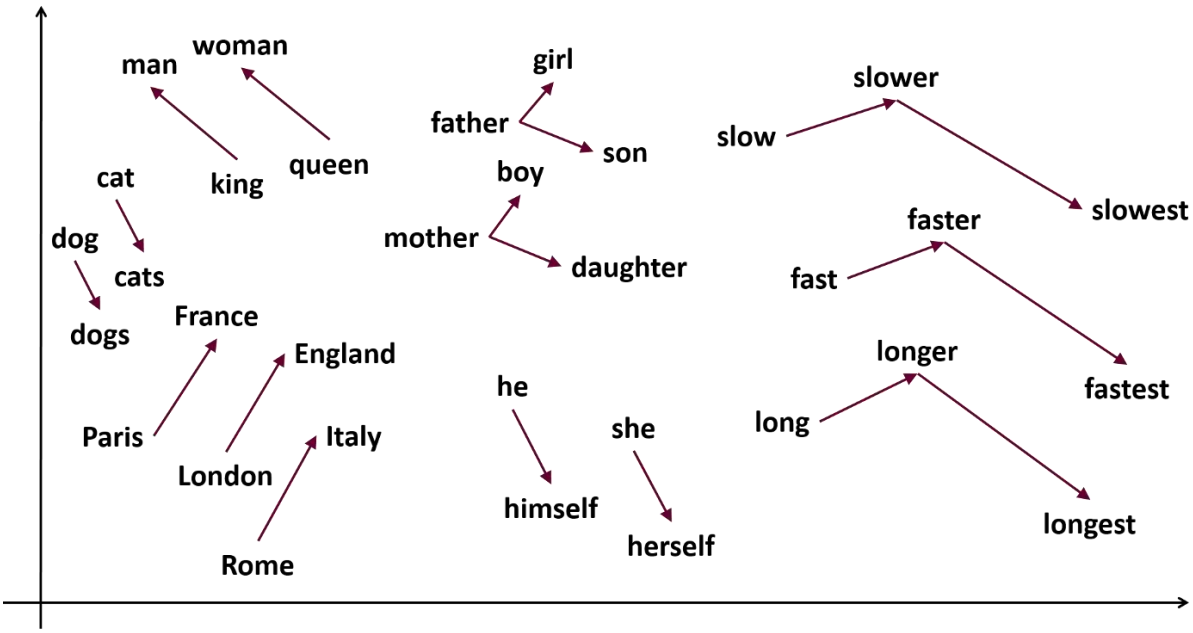


Figure 3.3: Illustration of semantic similarities of the word embeddings.[179]

Another intuition to the functioning of word embeddings is visualised in Figure 3.4 from Alammar.[180] Each word is represented by a dense vector of dimension 50, with colors ranging from red (value close to 2) over white (close to 0) to blue (close to -2). Similarities and differences can be observed when comparing the colors of the vectors between the different words. For instance, all words represent persons except the last row "water". This difference can be seen in the embeddings where the persons have a strong blue line in the center. Moreover, the embeddings for "girl" and "woman" are similar in many places.

Intuitively, *Word2Vec* aims to learn the representation of every word based on the other words surrounding it. Thus, the word embeddings are able to capture contextual relationships between words. For a more formal introduction into *Word2Vec* refer to the papers from Mikolov.[61,62]

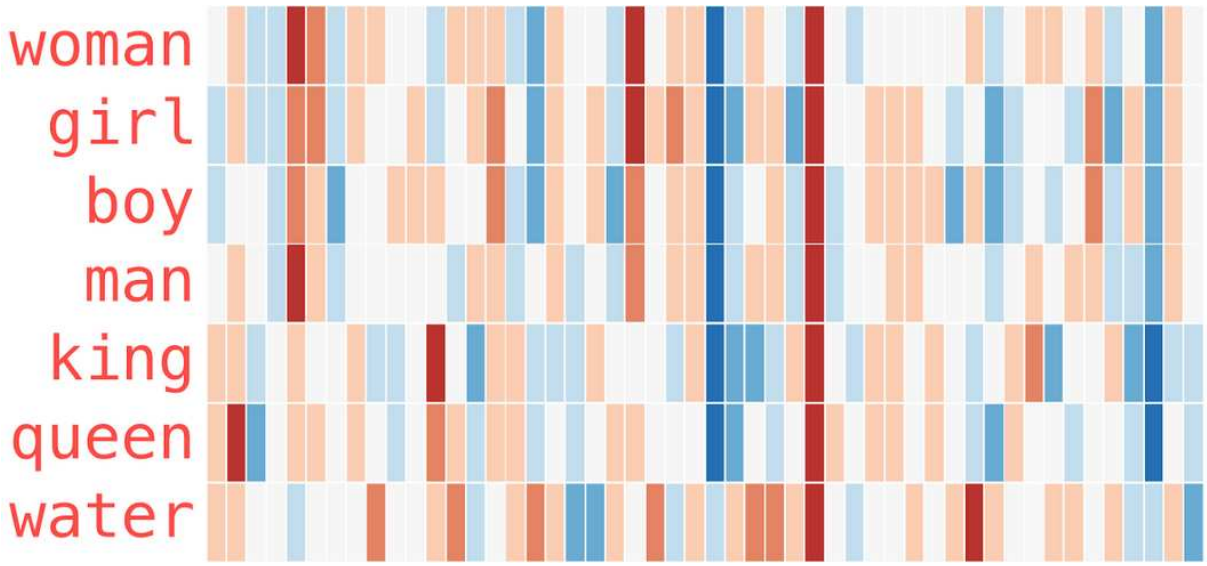


Figure 3.4: Word embeddings of 50 dimensions for several words. Color codes are based on their values, going from red (close to 2), over white (close to 0) to blue (close to -2).[180]

Some efforts have been made to extend the *word2vec* method. Le and Mikolov prolonged word embeddings to sentence and document embeddings.[181] Pennington et al. proposed *GloVe* (Global Vectors for Word Representation) leveraging the corpus statistics by calculating word embeddings on global word-word co-occurrence counts.[177] The *FastText* model extends *word2vec* by taking subword information into account.[63] In this method a vector is calculated not only for each word but also for the n-grams, allowing to approximate misspelled or out of vocabulary words. This is a practical property when working on noisy Twitter data and a reason why *FastText* word embeddings have been applied throughout this thesis.

Word embeddings are perfectly suited in a transfer learning setting, as they are typically pre-trained on a large corpus of text data to produce meaningful vectors in a first step. In a second step and given a specific task those word vectors can be used as features to train a machine learning model. However, it is important that the data of the pre-training and the data of the desired task resemble each other. Word embeddings trained on a general Wikipedia corpus won't be as impactful when

applied on Twitter data as the language used on Twitter differs quite significantly from the one on Wikipedia, which generally is free of misspellings, invented words, or grammatical mistakes.
Word embeddings are often used as a starting point for machine or deep learning algorithms and serve as input features.

The cosine is commonly used as distance measure in natural language processing to measure the similarity between two word vectors. Formally *cosine similarity* can be defined by the angle, or cosine of the angle, between two vectors and is given by:

$$sim(v_1, v_2) = \frac{v_1^T v_2}{||v_1||_2 \, ||v_2||_2}$$

, where $|| \cdot ||_2$denotes the euclidean distance and *sim(v1, v2)* ranges from -1, meaning an opposite similarity, to +1, stating an exact match of similarity with an angle of 0. And the *cosine distance* is defined by:

$$dist(v_1, v_2) = 1 - sim(v_1, v_2)$$

and only defined for positive values.

Throughout this work cosine similarity was applied to determine the similarity between words, sentences or tweets.

3.4.2 Contextualized Word embeddings

Word2vec, *Glove* or *FastText* word embeddings are context independent in a way that these models output one single word vector embedding for each word. In consequence, all different meanings of a word are encoded in a single vector. The vector embedding for the word “bank” in the sentences “*The money on my bank account*” and “*At the river bank*” are the same, neglecting the different word senses depending on the context.

The second revolution in NLP tackled this issue by providing a new generation of context-sensitive word embeddings. This profoundly moved the NLP field onto the next level, being able to calculate different word embeddings for “bank” depending on its context. In 2018, several approaches have been proposed to calculate deep contextualized word embeddings based on language modeling, such as *ELMo*,[182] *ULMFiT*,[183] *GPT*[184] or *BERT*.[64] Given the relevance for this work, only the *BERT* model,

probably the most utilized one, will be introduced in the following. *BERT* is designed on central elements for modern natural language processing: the attention mechanism and the transformer architecture.

Transformers

Before the emergence of the *Transformer* architecture,[65] recurrent neural networks[185] and in particular, one of its variants the Long-short-term-memory (LSTM)[186] dominated sequence-to-sequence tasks, such as machine translation or text summarisation. However, their weakness of modeling long sequences and their sequential nature prevented efficient computation and led to the *attention mechanism*, a fundamental innovation in neural NLP and the opening for contextual word embeddings.

The attention mechanism was proposed by Bahdanau[187] and is a way of selectively weighting different words in the input such that they will have an impact on the hidden states resulting in a weighted context vector that weights the outputs of all previous prediction steps. In short, attention allows the model to focus on the relevant parts of the input sequence.

The transformer architecture exploits the attention mechanism while processing sequences in parallel. All words are handled simultaneously rather than word-by-word. This architecture is boosting the training time significantly through heavy parallelization. For a broader introduction in the attention mechanism and multi-head attention refer to ANNEX 2.

BERT

Bidirectional Encoder Representations from Transformers, *BERT* in short, is a masked language model relying on a multi-layer bidirectional transformer encoder architecture.[64] It was introduced in two versions: $BERT_{base}$ which stacks 12 transformer encoder layers with 110 million parameters and a vector dimension of 768; and $BERT_{large}$ which stacks 24 transformer encoders with 340 million parameters and a vector dimension of 1024. Two training strategies are performed: Mask Language Model (MLM) and Next sentence prediction (NSP). In the MLM some input tokens are masked randomly, meaning replaced by the token [MASK], and the model tries to predict the masked token based on its context. In the NSP task, given a pair of sentences the model tries to identify if the second one follows the first sentence with the aim of capturing more long-term information. Training *BERT* on both training strategies results in a pre-trained model, which can then be fine-

tuned on various downstream tasks using labeled data. In this context, the transfer learning paradigm is adopted. Typically, the *BERT* model is pre-trained on a large dataset, and fine-tuned on a small task-specific dataset. With this training system *BERT* outperformed many task-specific architectures and achieved state-of-the-art results on a wide range of NLP tasks.[64] The input for the *BERT* model is a sequence of tokens. A *BERT*-specific tokenizer then adds two artificial tokens to each sequence. The first is the [CLS] token denoting the first token of the input sequence and can be used for classification tasks as it unites sentence specific information. The second token is [SEP] added between two sentences and is used in the NSP task during pre-training. *BERT* produces a 768 dimensional vector in its base version.

Various extensions of the *BERT* model have been proposed with the aim of optimising the model and adapting it to specialised use cases such as: *RoBERTa* (Robustly optimised *BERT* pre-training approach) which applies various modifications in the training step of the *BERT* model such as removing the next sentence prediction task or using longer sequences in the input data during training;[188] *DistilBERT* used distillation to reduce the *BERT's* model size by 40% and being 60% faster while retaining almost entirely its language understanding capabilities;[189] *BioBERT*, a pre-trained biomedical language model for biomedical text mining;[190] *BERTweet*, a pre-trained language model for English Tweets.[78]
In contrast to *word2vec* or *FastText*, which can be trained on a large dataset using multicore CPUs, *BERT* models rely to a great extent on training on GPUs.

BERT-based architectures, and in particular *BERTweet,* were examined in the second objective in both the preprocessing and causal models' creation part.

3.5 Supervised algorithms

Machine learning algorithms are all about pattern recognition in data. They are a useful tool to distinguish relevant data from non-relevant data. Supervised learning methods learn to perform predictions on unseen data based on human-labeled training data. In this way, the classifier acts as a mapper between the input data and the labels.

3.5.1 Support vector machine classification

In the first and second objective of this thesis, the supervised machine learning algorithm Support vector machine (SVM) was used in the preprocessing pipeline to filter only relevant tweets and in consequence remove noise.

The SVM algorithm aims to find an optimal hyperplane that best separates the classes by maximising the distance between the hyperplane itself and the closest data points, so called *support vectors.*[67,191] The hyperplane is defined by the linear function $\omega^T x + b = 0$, where ω is the normal vector to the hyperplane, x the data. Contrary to many other machine learning algorithms which output probabilities, the SVM provides the class identity: if $\omega^T x + b > 0$, SVM predicts the positive class and for $\omega^T x + b < 0$, the negative one, compare Figure 3.5.

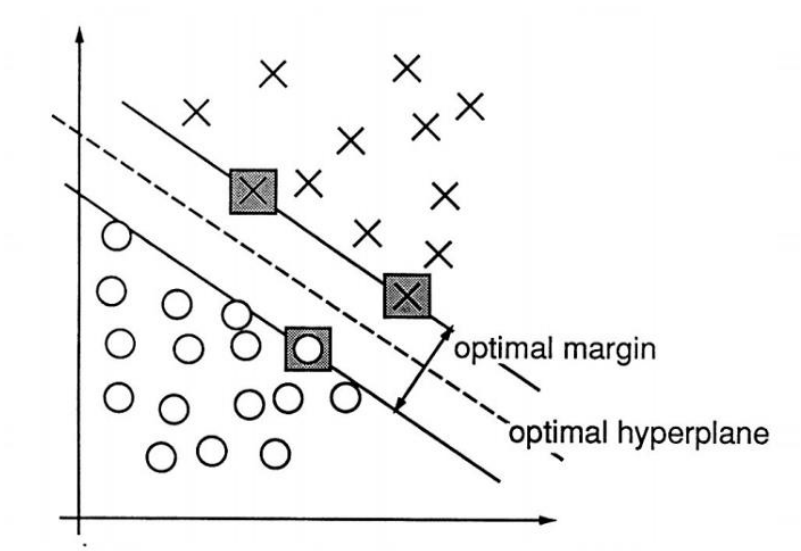


Figure 3.5: Support vectors, marked with grey squares, define the margin of the largest separation between the two classes.[191]

Additionally to undertaking linear classification, SVMs can also perform non-linear classification exploiting the so-called *kernel trick*. In situations when the data can not be separated linearly, the *kernel* function maps the input data to a higher dimensional space in which the data can be separated by a linear hyperplane such that the distance between the hyperplane and the data is maximal. The hyperplane is optimal if the margin is maximised with respect to the training data.

The overall idea can be described as: with a high enough number of dimensions, it is always possible to find a hyperplane separating two classes.

In the first objective the SVM was integrated in the preprocessing pipeline and in the third objective the SVM comprised the internal classifiers in the *classifier nodes*.

3.6 Unsupervised algorithms

Unsupervised algorithms, also frequently referred to as clustering algorithms, aim to find groups of similar documents in a collection of documents.

3.6.1 *K-means* clustering

The *K-means* clustering algorithm is a simple and widely used unsupervised algorithm creating a "flat" decomposition of data into a set of clusters.[68] The first step in the algorithm is to randomly select *K* data points as initial cluster centers (centroids). In the second step, each data point will be assigned to its nearest centroid based on a distance function *d*. In the third step, the centroids of each cluster are recalculated as the mean of all data points in the cluster. Then repeat step 2 and 3 until convergence is achieved and clusters are stable. Cosine similarity is usually taken as a distance function when working on text data. One of its challenges is to specify the number of clusters *K* to be identified by the algorithm.

The first objective leveraged the *K-means* algorithm to detect clusters from diabetes-related tweets.

3.6.2 Hierarchical clustering

Hierarchical clustering (HC) is a form of clustering in which the solution is presented in the form of trees. Different levels in the tree correspond to different levels of abstractions of the data. Hierarchical clustering is ideally situated for interactive exploration and visualisation as it enables the extraction of various flat partitions of clustering solutions of different levels of granularity.[69] Moreover, it may be more natural to reveal the underlying structure of the data in a hierarchical rather than a flat style.[192] HC can be broadly divided into *agglomerative (bottom-up)* and *division (top-down)* clustering. In the more common one, *agglomerative HC,* each data point is assigned to its own cluster and then pairs of clusters are merged repeatedly until a single root cluster is obtained; whereas in *divisive HC* initially all elements start in the same cluster and are then consecutively split.[193]

A novel hierarchical clustering algorithm was developed in the third objective building the foundation of the clinical decision support tool.

3.7 Named entity recognition

Named entity recognition (NER) is a natural language task aiming to identify proper entities and classify them into predefined categories such as dates, products, names or in the biomedical domain diseases, risk factors, proteins or mutations. Commonly the task of NER is formalised as a sequence classification task. A sequence labeling model is trained to assign a tag class to each token or unit within a sequence of tokens.

A standard method to identify entities in a text are conditional random fields (CRFs), a statistical sequential classification method.[70] A CRF is a form of undirected graph that defines a single log-linear distribution over label sequences given a particular observation sequence. CRF models the label sequence jointly instead of decoding each label independently by considering the correlations between labels in neighborhoods and jointly decoding the best chain of labels for a given input sentence. For a sentence with words $x_1, \ldots, x_T$ and associated tags (or hidden states) $y_1, \ldots, y_T$, a CRF is given by:

$$p(y \mid x, \theta) = \frac{1}{Z(x)} exp \left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_t , y_{t-1} , x_t) \right\}$$

The first term $Z(x)$ in the equation is a normalisation of all possible state sequences in order to output a probability:

$$Z(x) = \sum_{y \in Y} exp \left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_t , y_{t-1} , x_t) \right\}$$

θ_k is the weight for a feature k with feature function f_k and is learned during training. Considering only the special case of linear-chain CRF, the feature function f_k looks at a pair of adjacent states y_{t-1}, y_t, the input sequence x and the current position in the sequence t.

The feature functions are the core component of CRFs and can be defined in various ways. For instance, a simple feature function might be:

$$f_1(y_t, y_{t-1}, x_t) = 1 \quad \text{if } y_t = LOCATION \text{ and } x_t = Paris; \quad 0 \text{ otherwise}$$

or if a word is capitalized:

$$f_2(y_t, y_{t-1}, x_t) = 1 \quad \text{if } x_t = capitalized; \quad 0 \text{ otherwise}$$

Frequently used features are Part-of-Speechs tags, whether the word is a digit, certain prefixes or suffixes. Moreover, word embeddings can be used as features, as will be shown in this work. An extensive overview over conditional random fields is provided by Sutton and McCallum.[194]

CRFs have been applied widely in NER detection systems such as drugs[195] and also on other language as English such as Czech[196] or morphologically rich languages like Turkish.[197]
It has also been shown that combining word embeddings with LSTMs and CRFs have been powerful models to identify entities[198] or simply use word embeddings as features for a linear-chain CRF.[199] Lample et al. examined bidirectional LSTMs combined with CRFs and features based on character-based and unsupervised word representations.[200] Recently, large scale language model pretraining methods such as the aforementioned BERT further enhanced the NER performance.[64]

Typically, NER tasks label tokens in the BIO (Beginning, Inside, Outside) tagging format introduced by Ramshaw and Marcus.[201] Each token is encoded in one of the following tags: "B-*" indicates the beginning of an entity, "I-*" inside the entity and "O" represents outside the entity. A simplified variant of this format is IO tagging relying solely on the "I-*" and "O" tag which can lead to superior results compared with BIO tagging.[202]
In the second thesis objective the two entities *cause* and *effect* were predicted applying the IO tagging scheme resulting in the following tags: "I-C" inside cause, "I-E" inside effect and "O" neither cause nor effect. For instance, the sentence *Prediabetes forces me to change my lifestyle* with cause *Prediabetes* and effect *change my lifestyle* was encoded as:

Sentence:	*Prediabetes,*	*forces,*	*me,*	*to,*	*change,*	*my,*	*lifestyle*
IO tags:	*I-C*	*O*	*O*	*O*	*I-E*	*I-E*	*I-E*

The second objective framed the cause-effect pair identification as a named entity recognition task.

3.8 Active learning

Manual annotation of data is essential for the successful development and evaluation of sophisticated machine learning systems. At the same time, it is time-consuming, highly expensive and thus remains challenging for research groups.[71] In addition, manual annotation can be subjective if clear prior rules are not defined and in consequence is a potential source of bias. Active learning (AL) is a sample selection method with the aim to minimize annotation cost while maximizing the performance of machine learning models by selecting training data in a *smart* way.[72] An active learning strategy tries to select the most informative data instances from an unlabeled dataset to be labeled by an *oracle*, in most cases a human. This concept is attractive in scenarios where unlabeled data instances are highly available, but obtaining labels is expensive. The extensive survey of Settles details several strategies to evaluate the informativeness of unlabeled data and how to choose training data.[72] The *uncertainty sampling* strategy is the simplest and most commonly used one, in which the instance is chosen for which the system is the least certain about how to label.[73] In the case of a binary probabilistic classifier, *uncertainty sampling* chooses the instances for which the system predicted probabilities near to 0.5. A more theoretically-motivated and less often used strategy is the *query-by-committee*[203] in which a committee of models with different hypotheses are used and and return the most informative query as the query the models disagree the most about how to label. In the *expected gradient length* strategy, the model selects the instance that would impart the greatest change to the current model if we knew its label.[204]

Active learning has been widely applied on text data and clinical natural language processing.[71,84,205,206] It has also been shown that using modern word embeddings (*Word2Vec*, *FastText*, *BERT*) achieve substantial improvement over more commonly used vector representations such as Bag-of-Words.[207]

In both the second objective and third objective an active learning strategy was adopted to efficiently increase the training data and increase performance while minimizing the annotation effort.

3.9 Evaluation metrics

3.9.1 Supervised metrics

Typically, supervised machine learning algorithms are evaluated on several evaluation metrics, commonly based on counting the number of correct and incorrect predictions of the model in comparison with some ground truth labels. The most important metrics used throughout this work are detailed below.

3.9.1.1 Binary classification

Table 3.3 provides a confusion matrix for a binary classification in which the model prediction and actual true labels are compared to extract the evaluation metrics. *True positives (TP)* are defined as the number of correctly identified instances labelled as belonging to the positive class. *True negatives (TN)* correspond to the correctly identified instances labelled as belonging to the negative class. *False positives (FP)* refers to the number of incorrectly identified instances belonging to the positive class and *False negatives (FN)* specifies the number of instances incorrectly identified as belonging to the negative class.

		Actual class	
		Positive (P)	Negative (N)
Prediction class	Positive (PP)	**True positive (TP)** hit	**False positive (FP)** type I error, false alarm
	Negative (PN)	**False negative (FN)** type II error, miss	**True negative (TN)** correct rejection

Table 3.3: Confusion matrix for binary classification.

Using these statistics, it is possible to define the following evaluation metrics to determine a model's performance:[208]

- **Accuracy:** tests how well a model correctly identifies or excludes a condition. However, accuracy can be misleading with high scores in scenarios of large class imbalance.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision:** fraction of detections reported by the model that were correctly identified as positive class. Another term for this measure is *Positive predictive value*.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** fraction of true conditions that the model was able to detect. Another term for this measure is *True positive rate* or *sensitivity*

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score**: harmonic mean of precision and recall

$$F1 = 2\,\frac{Precision * Recall}{Precision + Recall}$$

In all objectives these performance measures served to evaluate the models.

3.9.1.2 Multi-class classification

In scenarios when the number of classes exceeds two, one refers to *multi-class classification*. This setting requires a redefinition of the above introduced evaluation metrics.[208] Table 3.4 illustrates the confusion matrix for the three classes *A, B, C* with class *B* being the reference class. For instance, if the actual class is B, but the model predicted class A or class C, then both are counted as *FP* due to misclassification.

		Actual class		
		Class A	Class B	Class C
Prediction class	Class A	**TN**	**FP**	**TN**
	Class B	**FN**	**TP**	**FN**
	Class C	**TN**	**FP**	**TN**

Table 3.4: Confusion matrix for multi-class classification. Here, for the classes A, B, C. Class B is the reference

A confusion matrix is built for each class being a reference class. Then there exist several possibilities to aggregate these scores: *micro, macro* or *weighted*. We only present the *macro* score which was applied in the second objective to evaluate the multi-class named entity recognition model.

- **macro:** calculates the metrics for each class and averages their values

  $Precision_c = P_c = \frac{TP_c}{TP_c+FP_c}$; $Recall_c = R_c = \frac{TP_c}{TP_c+FN_c}$;

  $macroPrecision = P_{macro} = \frac{1}{C}\sum_{c=1}^{C} P_c$; $macroRecall = R_{macro} = \frac{1}{C}\sum_{c=1}^{C} R_C$

The F-score is estimated similarly to the definition in the binary case by replacing precision and recall with *macroPrecision* and *macroRecall*.

3.9.2 Unsupervised metrics

3.9.2.1 Silhouette score

The *Silhouette* score helps to evaluate the correctness of the assignment of a data point to a specific cluster instead of another cluster by measuring both intra-cluster distance a and inter-cluster separation b.[209]

For a data instance i in the cluster C_i , the average distance between i and all other data instances in the cluster C_i is defined as:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

with $d(i,j)$ denoting a distance function.

The smallest average distance of a data instance i to all points in any other cluster is defined as:

$$b(i) = min_{i \neq k} \left\{ \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \right\}$$

The *silhouette score* of a data point i is then defined as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \in [-1, 1] \quad \text{if } |C_i| > 1$$

and

$$s(i) = 0 \quad \text{if } |C_i| = 1.$$

A value close to +1 indicates that the data point is far from the other clusters. A value close to 0 specifies that the data point is very close to the decision boundary between two neighboring clusters. A negative score indicates that the sample might have been associated with the wrong cluster. A global score can be obtained by averaging the *silhouette score* over all samples.

The Silhouette score was used in the first objective to determine the optimal number of clusters for the clustering algorithm *K-means*.

3.9.2.2 Hierarchical clustering

A global measure taking into account the overall set of clusters that are represented in a hierarchical clustering tree is the *F1-Score* as introduced by Larsen and Aone and applied by Zhao and Karypis.[69,210]

Each document is associated to a class c among all classes C and in an ideal clustering a cluster (node) k regroups only documents of a respective class c.

Let's define n_c as the total number of documents for a class c. For a cluster k among all clusters K, the total number of documents is n_k. The number of documents of class c in cluster k is then described by n_{ck}. The precision, recall and *F1-Score* for a given class c and cluster k are defined as

$$P(c,k) = \frac{n_{ck}}{n_k}$$

$$R(c,k) = \frac{n_{ck}}{n_c}$$

$$F(c,k) = \frac{2 * P(c,k) * R(c,k)}{P(c,k) + R(c,k)}$$

The *F1-Score* of the entire class c is the maximum *F1-Score* of any cluster/nodes in the tree,

$$F1(c) = max_k F(c,k)$$

leading to a global *F1-Score* as the weighted sum of individual class *F1-Scores*

$$F1\text{-}Score = \sum_{c=1}^{|C|} \frac{n_c}{n} F1(c)$$

where n is the total number of documents.

In an optimal hierarchical clustering, every class would have a corresponding cluster containing the exact same documents, leading to *F1-Score = 1*.

This $F1$-*Score* was applied in the third objective to evaluate the hierarchical clustering part of the clinical decision support tool.

3.10 Software

The principal programming language used to realise this work was Python, especially in the first two objectives with Scikit-learn, PyTorch and Gensim as main packages.[83,211–213] JavaScript with its visualisation library D3 was used in the second objective as well.[214] The third objective was developed in Scala based on the framework Apache Spark and Apache Lucene.[215–217] JavaScript and D3 were again applied for the visualisation and interface.

All the algorithms developed in this thesis are open-source under the GitHub account of the World Diabetes Distress Study: https://github.com/WDDS/ and Epiconcept's Sparkly package: https://github.com/Epiconcept-Paris/sparkly.

CHAPTER IV: IDENTIFICATION OF DIABETES AND DIABETES DISTRESS PROFILES

This chapter will outline the results pertaining the first objective of this thesis of identifying diabetes and diabetes distress related topics from social media data.

The content of this chapter has been peer-reviewed and published in the international journal: BMJ Open Diabetes Research and Care (2020).[79]

4.1 Introduction

Currently, the sources of stress, anxiety and concerns among people with diabetes are still underexplored and it is challenging for existing evaluation scales (e.g. PAID, DDS) to capture them, due to a lack of completeness (e.g. fear of hypoglycaemia, feelings of powerlessness or work-related stress). [9–12,218]

Social media data offer a unique opportunity to enlarge our understanding of diabetes-related distress and assess sentiments of people with diabetes with the help of the active online diabetes community, in particular on Twitter.[219]

Besides, exploring psychological information related to diabetes and associations between socio-economic factors and diabetes-related concerns based on real-life social media data is still a neglected area. Diving into social media and exploiting its data may provide new perspectives and enable future interventions to be more appropriately tailored concerning prevention, management and treatment of diabetes.

The objective of this study was the identification of diabetes-related concerns and the detection of associated emotions and sentiments centered on Twitter data from the USA.

4.2 Methods

In this study a strict preprocessing pipeline was applied to remove noisy and irrelevant tweets, which will be detailed in the following paragraphs. Figure 4.1 provides an overview over the preprocessing steps, followed by the identification of diabetes-related topics of interest and concluded by associating these topics with socio-economic factors such as the mean household income.

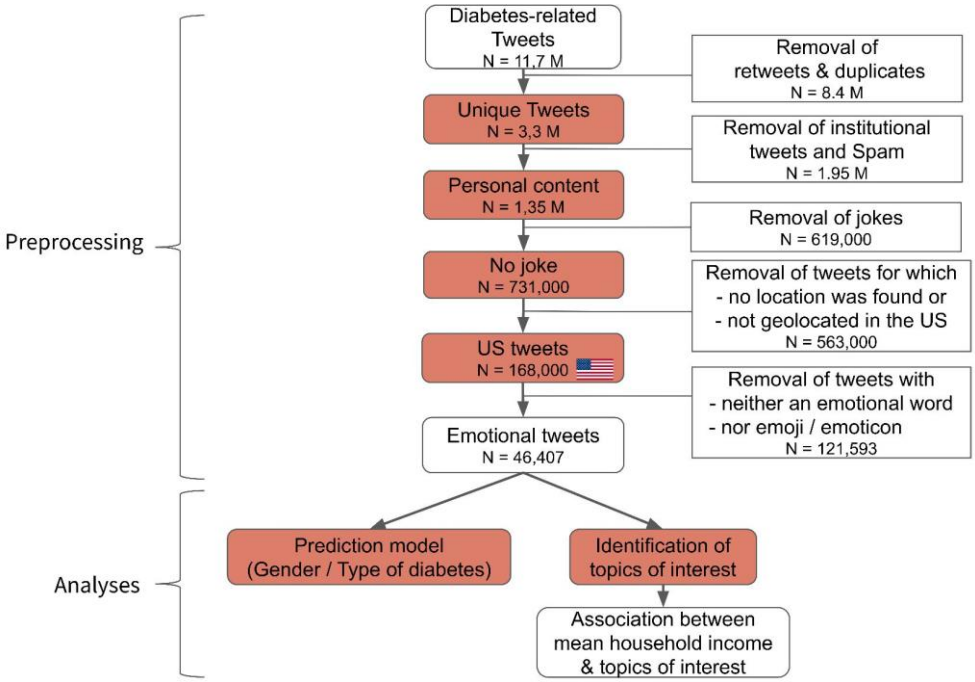


Figure 4.1: Workflow for the identification of diabetes-related topics.[79] The red steps involved machine learning steps

4.2.1 Data

This study was based on 11.7 million diabetes-related tweets from all over the world, in English language, collected from April 2017 to July 2019.

4.2.2 Data preprocessing

Data preprocessing is key in addressing the noisy and unstructured nature of Twitter data. The following sections outline each step that has been conducted.

4.2.2.1 Task-specific word embeddings

Word embeddings have been chosen to translate the textual data into a numerical format which is understandable for machine learning algorithms. Choosing suitable word embeddings is critical to ensure a healthy model performance. It has been shown that domain-specific input corpora are better in extracting meaningful semantic relationships than generic pretrained *Word2vec* or *GloVe* embeddings.[220] Following this advice we trained our own word embeddings using the *FastText* algorithm[63] on the initial 11.7 M tweets of our target domain diabetes. The *FastText* implementation from the topic modeling package *Gensim* was utilized to calculate the word embeddings.[213] In experimental tests we found that the best configuration was a vector dimension of 300, window size of 5, n-grams from 3 to 6, min count of 1, 20 training iterations and no preprocessing of the tweets. A vector representation of the whole tweet is then modeled as the average of its word vector embeddings.

4.2.2.2 Unique database

To guarantee a clean database consisting of only unique tweets, retweets and duplicates were removed. Moreover, we observed that certain Twitter accounts, probably chatbots, share the same message repeatedly with only slight modification, for instance changing a word. These accounts could potentially be chatbots which are becoming more and more popular.[221] Those slight changes prevented them to be recognised as duplicates and in consequence to be removed from the database of unique tweets. To tackle this issue and remove those almost identical tweets, cosine similarity was applied to each tweet pair based on *FastText* embeddings. If the cosine similarity exceeded 0.98, they were recognised as identical tweets and in consequence the duplicate was deleted. This threshold was determined experimentally.

Furthermore, we hypothesized that user mentions and urls create more noise than providing useful information for our analyzes. Thus, user mentions starting with "@" were replaced with the constant value "USER" and urls starting with "http…" were replaced with the constant value "URL".

4.2.2.3 Personal content classifier

Coming from a public health perspective, we were interested in analyzing tweets with *personal* content shared by people with diabetes or talking about diabetes. As *personal* content we understood information shared about their life, their opinions, their feelings and concerns.

In consequence, *institutional* tweets regrouping general, not to once personal life specific information such as health information, news articles or commercial tweets. Table 4.1 provides some examples of personal and institutional tweets.

Johnsen et al. tried to identify personal content based on personal pronouns like 'I', 'me', 'us', etc.[222] Contrary to this approach, a supervised strategy was adopted to identify tweets with personal content in a two step process.

Tweet	Personal (1=yes; 0=no)
Seven yoga poses for diabetes URL	0
if i go too easy on insulin my blood sugar spikes I risk later life problems. too much insulin and i go low and HAVE to eat or go into a coma	1
treating a hypo at 11:30 pm sucks ass he went from 225 to 64 then 56 🙄 please let this juice work its magic #type1diabetes	1
The latest Diabetes Daily! URL #t1d #dblog!!	0
Research links chemicals found in plastic used for food packaging to #type2diabetes in men: URL	0

Table 4.1: Examples of personal tweets and institutional tweets. URLs are replaced by "URL". Tweets were slightly altered to ensure privacy.

First, a machine learning algorithm was trained to detect tweets from personal users. For this purpose, 2275 randomly chosen tweets were manually labeled for being *personal* (label: 1) or *institutional* (label: 0). Two researchers labeled the first 300 tweets, according to the definition at the beginning of the paragraph, discussed disagreements and one researcher continued to label the other tweets. The input for the machine learning algorithm was a tweet embedding obtained by averaging the word vector embeddings of all words in the tweet based on the trained diabetes *FastText* embeddings. The trained algorithm was then applied on all tweets per user and if their average prediction was above 0.25 the tweets were categorised as coming from a personal user. This threshold was determined experimentally. For example, if in our entire dataset a user X shared 5 tweets with predictions of "0, 0, 0, 1, 0", then the user was considered an institutional user, as the average prediction was 0.2, and consequently all his/her tweets were removed from the dataset. This first classifier acts as a crude filter, removing obvious *institutional* accounts.

Secondly, a personal user can tweet both tweets with personal but also institutional content. To separate these two categories, another 1884 tweets were randomly selected from all tweets of

personal users and manually labeled according to personal content or institutional content. Identically to the previous annotation, two researchers labeled the first 300 tweets, discussed disagreements and one researcher continued to label the other tweets. Analogous to the first classifier, a second machine learning classifier was trained on those tweets and then applied on all personal user tweets to obtain a database consisting of tweets from personal users with personal content.

Several machine learning algorithms were tested to determine the optimal performance for the personal classifiers: Support Vector Machines, Logistic regression and Random Forests.

4.2.2.4 Joke classifier

Jokes about diabetes are a common phenomenon on Twitter and a large part of jokes are related to food such as "*I ate so much ice cream, that it will give me diabetes haha"* or "*I can taste the diabetes in my tea*" (fictional tweets). In our study, jokes are considered noise and misleading and in consequence were removed.

Automatic detection of jokes, humour and irony in a piece of text is a challenge, even for humans. Contrary to Zhang and Liu who identified humorous tweets based on phonetic, morpho-syntactic, lexico-semantic, pragmatic and affective features, we again leveraged in this work a word embeddings in combination with a binary classifier.[223] For this purpose 998 random tweets were labeled in jokes (n=250) and non-jokes (n=748). Again, two researchers labeled the first 200 tweets, discussed disagreements and one researcher continued to label the other tweets. Due to the class imbalance, the data augmentation technique Synthetic Minority Oversampling Technique (SMOTE) oversampling was implemented, using the package *Imbalanced-learn.*[224,225] In contrast to random oversampling where the data instances of the minority class are simply duplicated and so do not add any new information to the model, SMOTE creates new synthetic data instances from the existing training examples which increases the variances in the training set. In detail, a random training instance *A* of the minority class is chosen and then its *k* (default: 5) nearest neighbors are found. One of the *k* nearest neighbors *B* is chosen randomly. The synthetic new instance is then generated by selecting a random point between the two points *A* and *B* in the feature space.

Analogous to the *personal* classifiers, several algorithms were tested: Support Vector Machines, Random Forests, XGBoost and Logistic Regression.

4.2.2.5 Geolocation

This study was conducted only on tweets from the United States of America. Yet, only a small fraction of users share their geo-location information (geo-tags). Some studies report that less than 3% of all tweets[149] contain geo-tagged information, while others report that less than 1% of tweets are geo-tagged and less than 50% of users provide valid city-level information.[226,227] Hence, identifying the geolocations of a tweet is a tough task. Sewalk et al. built a geolocation engine leveraging the tweets metadata and combined them with natural language processing methods and external services such as Google Maps Geocoding application programming interface to infer a location.[149] Another approach extends the *FastText* architecture to account for various user metadata.[228]

We also developed a geolocation engine to exploit the metadata of tweets and thus aim to geolocate as many tweets as possible. Similarly to Shah et al., who identified the location of a tweet based on the user-defined *user location* field in the user profile,[229] our geolocation engine exploits the *user location* field. But in addition, we also include the geo-tagged *place full name* and the *user description* field of the user's biography in this process. Two external components were necessary for the geolocation algorithm. First, the open-source geographical database *GeoNames* was downloaded, containing over 11 million geographical names worldwide.[230] Secondly, the open-source software Apache Lucene was utilized, which enables full text indexing and searching capability combined with Apache Spark to process the large amount of data in an efficient way.[216,217]

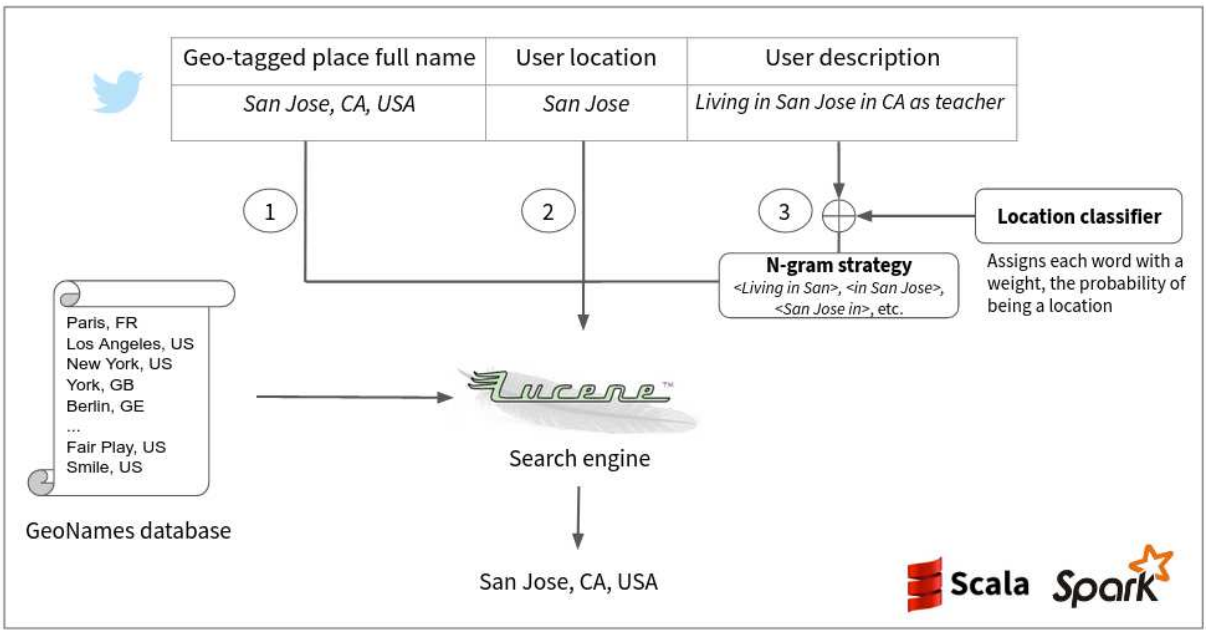


Figure 4.2: Geolocation of tweets using the metadata

Figure 4.2 demonstrates the workflow to identify locations. First, the geolocation engine checks the presence of the most reliable and exact indicator for a location, the '*place full name'* field in the meta-data. If present, it was used as the query. Otherwise, the second option to infer the location from was the *user location* field. If both fields were not available, we tried to extract the location from a third option, the *user description*, where users share a personal note such as their profession, what they like or where they live. The query is then matched with the GeoNames database, using Lucene, to extract the most similar location. Several configurations have been investigated to improve this matching performance. The first one added the Levenshtein distance as an additional parameter for the Lucene.[231] It describes the minimal number of single-character changes, such as insertions, deletions and substitutions) to change a word into another. A useful feature when working with noisy Twitter data, which links misspelled to correctly spelled words. For instance, the Levenshtein distance between "diabetes" and "diabete" is 1 (deletion) or between "Japan" and "Iapane" is two (substitution J -> I, insertions e). The motivation for another optimization was that Lucene's standard query concatenates all words with an OR operator, which performs well for small queries like the *user location* or *place full name* field, but lacks precision for longer texts. For the *user description* field the standard query produced a lot of false positives as many non-location words are present in the query which inconveniently often matched a location in the vast GeoNames database. For this reason, we developed the *N-gram* strategy for the user description query, with a default parameter of 3. For instance all 3-grams in the query "Living in San Jose in California as teacher" are <"Living in San">, <"in San Jose">, >"San Jose in">, <"Jose in California">, <"in California as">, <"California as teacher">. Instead of sending the whole query to Lucene, all 3-grams are sent separately to Lucene reducing the query size and thus the probability that an unimportant word gets matched with a location. An additional clue has been added to the N-Gram strategy to avoid misclassification and further improve performance. Taking the same example, imagine the highest matching score was returned for the N-gram "San Jose in" and the matched location returned from the database was San Jose in Costa Rica. To avoid these errors, the N-gram with highest scores, in the example "San Jose in", gets extended by a word to the left "in San Jose in" and to the right "San Jose in California" and both extensions are resent to Lucene with the aim of improving the score and prediction. Ideally the query "San Jose in California" scored higher and returned the correct match San Jose in California from the database. The extension by a word continues to either right or left until a maximum score in combination with the most confident location prediction is returned.

Working with textual data always bares surprises. Another subtlety that occurred was the fact that quite original location names exist in the real world. For instance, there are places called “Smile” or “Fair Play” and if a user has written “I like fair play” in his *user description*, then the algorithm would return the city “Fair Play”, which is obviously wrong in this context and a false positive. To adjust for these errors and avoid wrong classifications of *user descriptions* an additional binary SVM classifier, the *Location classifier,* was developed to decide if a word is location. The classifier was trained on 5000 randomly chosen US locations from the GeoNames database and 5000 randomly chosen words from the vocabulary of the trained diabetes *FastText* embeddings which also served as embeddings for the classifier. The *Location classifier* detects if a word is a location with an accuracy of 91.3%, a precision of 92.9%, recall of 89.5% and F1 score of 91.1%. The binary output of the SVM was then transformed into probabilities using an improved version of Platt’s scaling, essentially a logistic transformation.[232] Those probabilities corresponded to an additional weight factor per word for the Lucene query, to ultimately improve the location prediction. For instance, in the above-mentioned example “I like fair play”, the *Location classifier* would attribute a low weight to the words “fair” and “play” and in consequence Lucene would attribute a lower score for returning “Fair Play” as location.

This geolocation engine allowed us to target Tweets coming from the US.

4.2.3 Identifying emotions

A special focus of this work lies on diabetes distress and on psychological factors and emotions related to the day-to-day disease management. To point the attention towards diabetes distress, solely tweets containing an emotional element were considered. In this context, an emotional element is defined by the occurrence of either emotional words or emojis / emoticons. The work of the psychologist Parrot provided us with a list of over 100 emotions classified in a hierarchical structure with six primary emotions: *joy, love, surprise, sadness, anger,* and *fear* (full list available in ANNEX 3). Each primary emotion is further divided into sub-emotion categories which themselves are subdivided.[233] We combined Parrott’s emotions with emotional keywords present in the two most common questionnaires to assess psychological health in people with diabetes, precisely PAID and DDS. To capture as many emotional tweets as possible, this list was further extended by hypothesising that synonyms of these emotions should be included as well. The synonym identification was realized using the WordNet database.[234,235] Emoticons comprise a metacommunicative pictorial representation of a facial expression using punctuation marks and

letters such as ':-)'.[236] Whereas emojis are pictograms used in electronic messages like 😄, 😍 or 😜. Wolny's classification of emojis and emoticons allowed us to map emojis and emoticons into emotional categories similar to the six primary emotions defined by Parrot.[237]

4.2.4 Sentiment analysis

Generally, sentiment analysis inspects the opinion expressed by users in their text communications, whether the opinion or attitude tends to be positive or negative. Sentiment analysis has been addressed on a wide range of health-related tweets. A review on sentiment analysis found that on average, 46% of health-based tweets contain some form of positive or negative sentiment.[238] Cole-Lewis et al. examined conversations about e-cigarettes;[239] Daniulaityte et al. studied supervised machine learning techniques in drug related tweets;[240] Hawkins et al. analyzed the feedback shared by patients based on their experience receiving healthcare in hospitals[241] and Adrover et al. focused on patient reactions and side effects to treatments.[242]

Sentiment analysis can broadly be divided into Lexicon based methods, machine learning methods or a mixture of both approaches. Lexicon based methods depend on a dictionary with positive and negative terms, such as Sentiwordnet.[243] The polarity of a text sequence is determined on the ground of the polarity of its words. However, their effectiveness strongly relies on the quality of the lexical resources and their performances are limited for their inability to account for contextual information, novel vocabulary or nuanced indicators of sentiment expression.[244] Machine learning approaches require a set of training instances, which may be costly to acquire. Also, they depend on the training set to represent as many features as possible, a bottleneck in sparse and short text of social media. Currently a lot of sentiment analysis tools exist, but until now none has emerged as gold standard method.[245]

In this work, we adopted the open-source and human-validated tool Valence Aware Dictionary for Sentiment Reasoning (VADER)[74] to compute the sentiments shared in tweets. It is a lexicon and rule-based sentiment analysis tool specifically developed for social media data and widely used.[149,245,246] It combines a lexicon and the processing of the sentence characteristics to determine sentence polarity. Moreover VADER includes grammatical and syntactic hints to communicate changes in the sentiment intensity such as treatments for negations (e.g. "not good"), punctuation (e.g. number of "!" in "Great!!!!!"), emoticons and emojis, capitalization (e.g. "I LIKE YOU") , degree modifiers to alter sentiment intensity (e.g. "Research is amazingly interesting" is more intense than "Research is

interesting"), constructive conjunctions to shift the polarity (e.g. "but") or understanding slang words (e.g. "sux") and acronyms (e.g. "lol"). VADER computes sentiment and valence for each word level and provides positive, negative, and neutral scores at the sentence level. Here, we used the compound score as our main metric for the sentiment analysis, also referred to as the SA score, which is a unidimensional and normalized measure of sentiment between −1 (most negative) and +1 (most positive).

4.2.5 Topic extraction

The core of this work is the identification of diabetes and diabetes distress profiles, as defined in section 1.2.6, and to characterise what people with or related to diabetes talk about. This task was framed as a clustering task to regroup tweets that are talking about similar topics. The clustering was applied on the personal, non-joke tweets containing an emotional element from the USA. The crucial property of word embeddings of words being similar in semantics are also similar in the vector space is again adopted for the clustering. We applied the unsupervised algorithm K-means with cosine similarity as distance measure to group the tweets into topics/clusters using the tweets embeddings, analogous to the previous classifiers. *K-means* was preferred over topic models (e.g. Latent Dirichlet Allocation) as it provides mutually exclusive groups and affects each tweet to a specific cluster, in addition to its capability of leveraging word embeddings. Thus it allows the study of associations with other factors such as sociodemographic factors. As for many clustering algorithms, the K-means algorithm requires the desired number of clusters to be obtained beforehand. Without having an assumption about the right number of clusters k, we exploited the Silhouette score to determine this parameter. Essentially, the Silhouette score measures how on average each data point is closer to its cluster's center than to any other cluster. We obtained the highest *Silhouette score* for $k = 30$. All tweets were then assigned to one of the 30 clusters based on the *K-means* algorithm. Each cluster is manually attributed with a label to characterise it better, according to the 10 most contributing tweets, the ones closest to the cluster center, and the most frequent words (top words) in the cluster. The term *topics of interest* is equally used for clusters.

4.2.6 Assessment of the mean income

An extension of this work was the investigation in crossing the identified topics of interest with external variables. We chose the socio-economic variable *household mean income* for the intersection with the identified topics of interest. Both the tweets were geolocalized on a city level and the data for the household mean income was extracted per city based on the 2017 American Community Survey from the US Census Bureau.[75] The mean household income was divided into tertiles: *low income* (US$ 24.609 - US$ 67.224); *medium (med) income* (US$ 67.225 - US$ 86.758) and *high income* (US$ 86.759 - US$ 394.259). Each tweet was then linked with the mean income for its geolocated city and assigned to its respective tertile *low, medium,* or *high income*. The mean income information in combination with the associated cluster of the tweet allowed us to calculate the associations between topics and mean income tertiles.

More formally, the probability of a tweet being of topic k, conditioned on being from *low, med* or *high* income, was calculated. For N being the total number of tweets, N_k being the number of tweets associated to topic k, N^{low} the number of tweets assigned to low income and N_k^{low} the number of tweets of topic k and of low income (equivalent definitions for *med* and *high* income), following conditional probability was evaluated:

$$P(topic(tweet) == k \mid income(tweet) == med) = \frac{N_k^{low}/N}{N^{low}/N} = \frac{N_k^{low}}{N^{low}}$$

$$P(topic(tweet) == k \mid income(tweet) == med) = \frac{N_k^{med}/N}{N^{med}/N} = \frac{N_k^{med}}{N^{med}}$$

$$P(topic(tweet) == k \mid income(tweet) == high) = \frac{N_k^{high}/N}{N^{high}/N} = \frac{N_k^{high}}{N^{high}}$$

For each topic the Chi-squared test was applied to calculate the *p-value* between the binary variable if a tweet belongs to the corresponding topic and the tertile categories of the city-level mean income.

4.2.7 Gender & Type of diabetes classifier

When working with textual data from Twitter, information about gender or the type of diabetes, a tweet refers to, rarely is available. Gender could be derived from the *user name*'s field in the metadata, however a lot of users do not provide their real name and rather prefer artificial names making it difficult to rely solely on this information. Nevertheless, we attempted to derive gender

and type of diabetes information from the tweet text and its metadata using machine learning algorithms.

For this purpose, a random subset of 1897 emotional tweets was manually annotated for gender (M: male; F: female; U: unknown) based on the *text, user description* and *user name* field and for type of diabetes (T1: type 1 diabetes; T2: type 2 diabetes; U: unknown) based on the *text* and *user description* field. Here, type of diabetes does not refer to a user's type of diabetes but rather if a tweet refers to a type of diabetes in a broader sense. For both classifiers a third class *Unknown* was added for cases in which gender or type of diabetes inference was not possible.

In the case of the gender classifier, the *user name* field is a strong indicator to derive the gender. We relied on this fact to augment the training data efficiently with the aim to enhance classification performance. Therefore, a list with popular baby names in the US from 2019, with information about the name, gender and occurrence, was downloaded from the U.S. Social Security Administration.[247] The most frequent names, with at least 500 occurrences, were chosen assuming that their word embeddings would be a more stable factor for recognising male or female names than the word embeddings of rarer names. This list of male and female names was matched with the first names of the *user name* field of 12.000 random tweets resulting in 3614 tweets having a male or female first name. Finally, these 3614 tweets were added to the original manually labeled 1897 tweets leading to a total of 5511 tweets to train the gender classifier.

For the model training of the gender classifier, the *FastText* embeddings for the tweet fields *text*, *user description* and *user name* were concatenated and in the case of the type of diabetes classifier the *FastText* embeddings of the fields *text* and *user description*. In view of our positive experiences with the SVM algorithm in combination with SMOTE oversampling for the minority class we kept this configuration for both the gender and type of diabetes classifier training.

4.2.8 Software

Throughout this project, Python version 3.6 was utilized with the main packages being *scikit-learn* v.0.23 (machine learning algorithms and data preprocessing methods), *gensim* v.3.8 (topic modeling and word representation) and *pandas v1.1.5*.[83,211,213,248] Algorithms related to this work were open-sourced under the following address:
https://github.com/WDDS/Tweet-Diabetes-Classification.

The geolocation algorithm was implemented in *Apache Spark* using the programming language *Scala* and *Apache Lucene* served as search and index machine. The geolocation algorithm is embedded in Epiconcept's *Sparkly* package (portable machine learning spark applications: https://github.com/Epiconcept-Paris/sparkly.

4.3 Results

4.3.1 Personal tweets

Model performance for both the personal user classifier and the personal tweet classifier were assessed by a 10-fold cross validation and a grid search was applied to determine the best model parameter configuration. Performances for the personal user classifier are summarized in Table 4.2.

Model	Acc	Prec	Rec	F1	Best model configuration
SVM	88.8	82.8	75.8	79.1	Regularization : 15.0; Kernel coefficient (gamma) : 0.01; Kernel: rbf; Tolerance for stopping criterion : 0.1
Logistic regression	88.4	84.0	72.1	77.6	Regularization : 0.5; Tolerance for stopping criterion : 0.01
Random forest	88.6	86.8	69.5	77.2	Max depth : 8 max features : 'auto' n estimators : 120

Table 4.2: Model performance for classifying personal users vs. institutional users for several machine learning algorithms (in %).

The accuracy of all three algorithms is quite close. However, the Random forest algorithm led to the highest precision with 86.8%, whereas the Support Vector Machines dominated the recall with 75.8% and F1-Score of 79.1%. With the aim of determining as many personal users as possible, we chose the algorithm with the highest recall, the SVM to be applied on all unique tweets. Besides, SVMs have been shown to achieve high accuracy on Twitter data.[249,250] Figure 4.3 shows the confusion matrices for the SVM classifier.

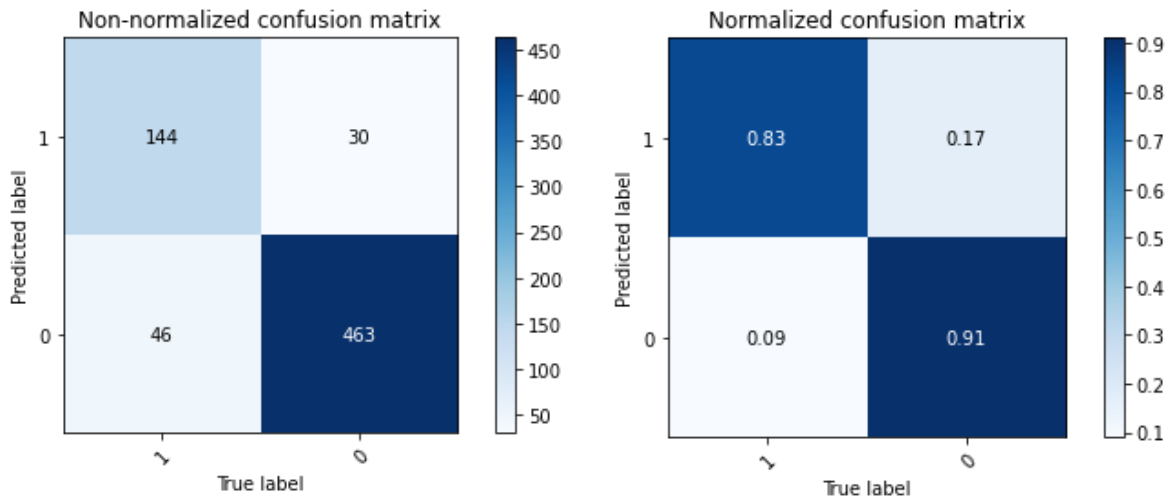


Figure 4.3: Confusion matrices for the personal user classifier (SVM) non-normalized and normalized

Concerning the personal tweet classification, all three algorithms performed similarly, especially the Support vector machine and Logistic regression which slightly outperformed the Random forest with a recall of 92.7% and F1-score of 92.0%, compare Table 4.3. To be consistent with the first personal user classifier and still with focus on the highest recall, we selected the support vector machine as well to classify personal tweets. The confusion matrix for the SVM is shown in Figure 4.4.

Model	Acc	Prec	Rec	F1	Best model configuration
SVM	92.6	91.4	92.7	92.0	Regularization : 15.0 Kernel coefficient (gamma) : 0.1 Kernel: poly Tolerance for stopping criterion : 0.1
Logistic regression	92.6	91.4	92.7	92.0	Regularization : 1.0 Tolerance for stopping criterion : 0.01
Random forest	92.2	90.7	92.7	91.7	Max depth : 8 max features : auto n estimators : 100

Table 4.3: Model performance for classifying personal tweets vs. institutional tweets for several machine learning algorithms (in %).

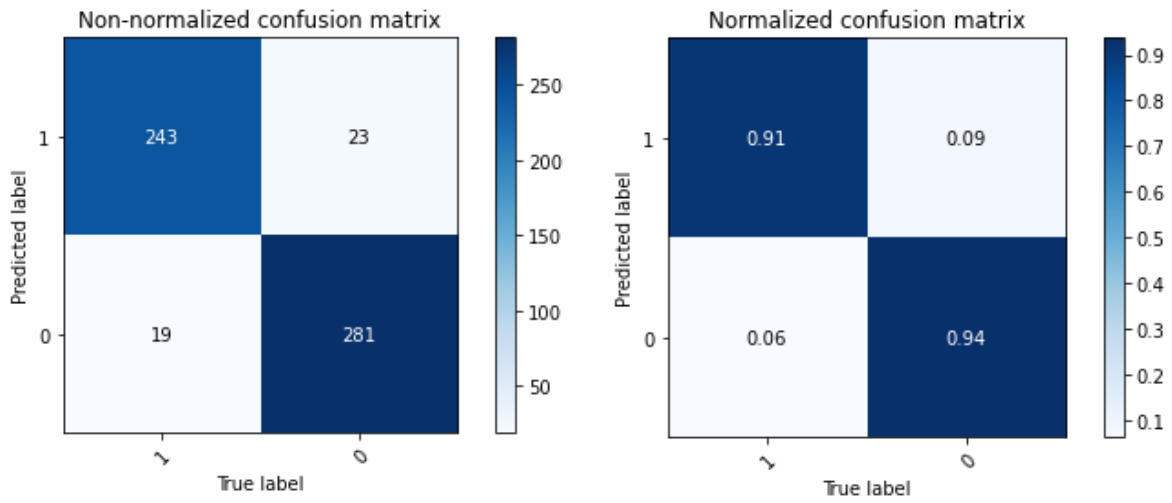


Figure 4.4: Confusion matrices for the personal tweet classifier (SVM) non-normalized and normalized

The personal user SVM classified all 3.3M unique tweets on a user level and only tweets from personal users whose average tweet predictions scored higher than 0.25 passed this barrier. The second personal tweet classifier was then applied on these tweets from personal users to identify tweets from personal users with personal content which led to a database of 1.35 M unique, personal tweets.

4.3.2 Joke classifier

A 10-fold cross validation was applied and the input for each tweet instance were the concatenated embeddings for the tweets' *text* and *user description* field. Again, a grid search allowed us to estimate the best model parameters including a weighting factor for the user description embedding. Besides, the joke classifier was evaluated on optimizing the recall to capture as many jokes as possible. This weight factor was multiplied to each value in the 300-dimensional word vector. This emphasized the importance of the tweet *text* field stronger.

Table 4.4 lists the performance over different classifiers with and without SMOTE oversampling. Generally, SMOTE oversampling improved the performance for all classifiers significantly. The SVM reached the highest recall with 82.7%, the random forest dominated the precision with 63.4% and the logistic regression achieved the highest accuracy of 80.3% and F1 68.1%. As we aimed to detect as many jokes as possible, the SVM with highest recall was prioritized and in consequence applied on all personal jokes. The removal of jokes in personal tweets, led to a dataset of 731.323 personal, non-joke tweets. Figure 4.5 shows the confusion matrix for the SVM classifier indicating a high

number of false positives.

Model		Acc	Prec	Rec	F1	Best configuration
SVM	SMOTE	76.7	54.5	**82.7**	65.7	Regularization : 0.1, Kernel : linear, Tolerance : 1.0, smote - k_neighbors: 5, weights : 'tweet': 1, 'userDesc': 0.2
	No SMOTE	77.3	58.0	58.0	58.0	
XGBoost	SMOTE	80.0	60.2	76.5	67.4	booster: gblinear, gamma: 0, learning rate: 0.1, max depth: 3, n estimators: 125, reg_alpha: 0, reg_lambda: 1.0, smote - k_neighbors: 6, weights: 'tweet': 1, 'userDesc': 0.2
	No SMOTE	73.0	0.0	0.0	0.0	
Random Forest	SMOTE	79.3	63.4	55.6	59.2	criterion: entropy, max depth: 5, max features: auto, n estimators: 150, smote - k_neighbors: 5, weights: 'tweet': 1, 'userDesc': 0
	No SMOTE	75.3	81.8	11.1	55.1	
Logistic regression	SMOTE	80.3	60.6	77.8	68.1	Regularization: 0.1, penalty: l2, tolerance: 0.01, smote - k neighbors: 6, weights: 'tweet': 1, userDesc: 0
	No SMOTE	75.3	54.3	54.3	54.3	

Table 4.4: Performance of the joke classifier in percentages over different classifiers. For each classifier performance is shown with and without SMOTE oversampling. The last column shows the parameters for the best model performance with SMOTE oversampling.

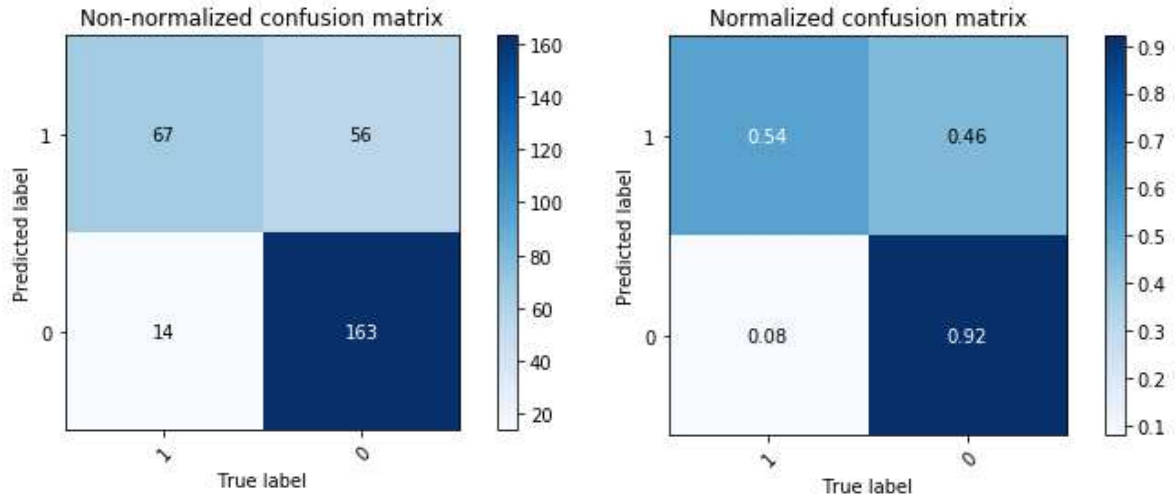


Figure 4.5: Confusion matrices for the joke classifier (SVM), non-normalized and normalized

4.3.3 Geolocation

The geolocation performances over the different configurations, to infer geolocated information on a city level, are summarized in Table 4.5.

The performance based on the *place full name* and *user location* field was evaluated on 236 manually labelled samples and achieved the best result of a precision of 85%, recall of 96% and F1-Score of 90% using the N-gram strategy and Levenshtein distance of 1. The influence of the *location classifier* on the geolocation prediction of these two fields was neglectable, due to the shorter text length and thus less irrelevant words for the geolocation. Whereas, the *location classifier,* in addition to the N-gram strategy and Levenshtein distance of 1, significantly boosted prediction performance on the *user description field*, which was tested on 214 manually labeled tweets and yielded a precision of 81%, recall of 69% and F1-Score of 74%.

Configuration	Acc	Prec	Rec	F1
Place full name / user location field				
Levenshtein distance = 0	72.8	72.7	100	84.2
Levenshtein distance = 1	78.8	78.1	99.4	87.5
Levenshtein distance = 1, Ngram strategy	**83.5**	**85.2**	**95.7**	**90.2**
Levenshtein distance = 1, Ngram strategy, Location classifier	83.5	82.7	99.5	90.3
User description				
Levenshtein = 0	56.5	25.5	51.1	34.0
Levenshtein = 1	70.6	49.0	78.7	60.4
Levenshtein distance = 1, Ngram strategy	80.8	68.0	76.8	72.1
Levenshtein distance = 1, Ngram strategy, Location classifier	**83.6**	**80.7**	**68.5**	**74.1**

Table 4.5 Geolocation algorithm performances for different configurations. On the top the performances are shown for the tweet fields: *place full name* and *user location*. On the bottom performance is shown for the field *user description*.

From the 731.323 personal, non-joke tweets, we found that only 6% (40.931/731.323) contained geo-coordinates shared from Twitter in the field *place full name*. In total, the geolocation engine was able to infer geo-coordinates for 63% of the tweets (463.623 / 731.323) worldwide, of which 31% (226.345 / 731.323) were found to be in the US and 23% (167.743 / 731.323) in the US that also had information on the city-level.

As a result, our work continued on the 167.743 tweets, containing geolocalized city information in the US.

The spatial distribution of these tweets is illustrated in Figure 4.6. California (N: 18.551) and Texas (N: 14.237) show the highest number of tweets, whereas Vermont (N: 197) and Wyoming (N: 131) had a low number of tweets. Those numbers correlate with the population distribution over the different states according to the US Census Bureau.[251] At the city level the dominant cities with the highest number of tweets are New York City (N: 9.663), Los Angeles (N: 5.301) and Chicago (N: 4.884).

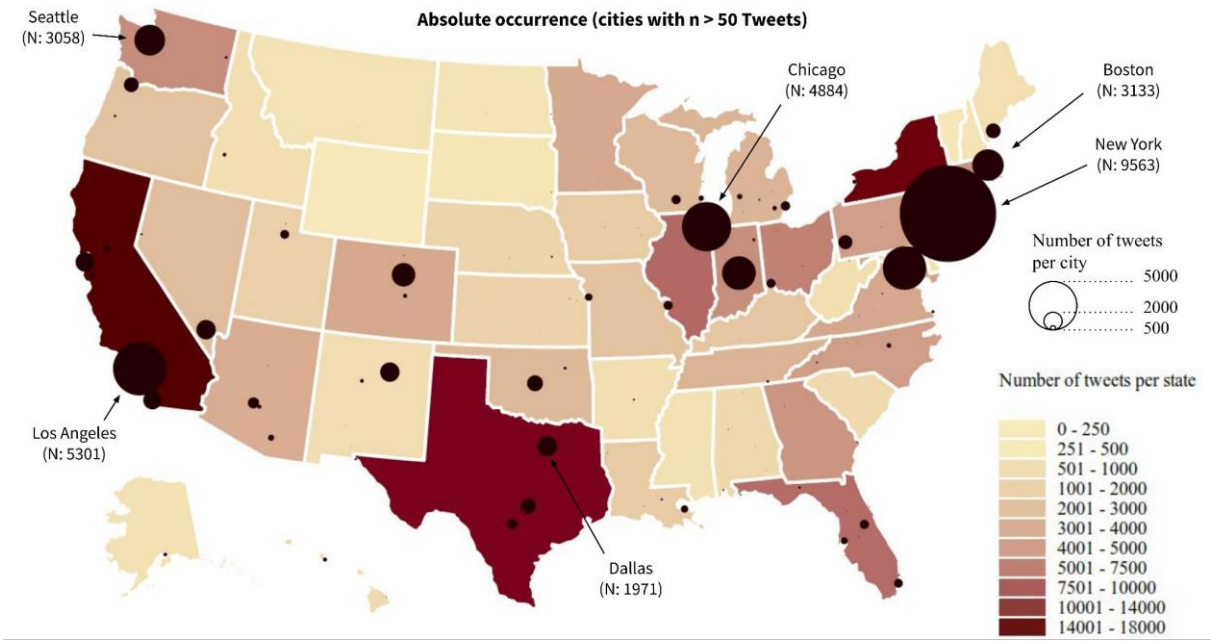


Figure 4.6: Spatial distribution of diabetes-related, personal, non-joke tweets over the USA. The number of tweets in each state is represented by color ranging from a few tweets in white-yellow to a lot of tweets in dark red. The size of the black circles is proportional to the number of tweets.[79]

4.3.4 Gender & Type of diabetes

A grid search and 10-fold cross validation was applied to train the Gender and Type of diabetes classifier. The accuracy for the Gender classification was 86% and more performance measures can be deducted from Table 4.6. With a precision of 92% a tweet could be identified as being from a man and with 93% for a woman. Figure 4.7 shows the confusion matrices in a non-normalized and normalized version.

	Prec	**Rec**	**F1**
M	0.92	0.87	0.89
F	0.93	0.89	0.91
U	0.36	0.62	0.46
macro avg	0.74	0.79	0.75

Table 4.6: Performances for the gender classifier (SVM) for the following classes: M = male; F=female; U=unknown

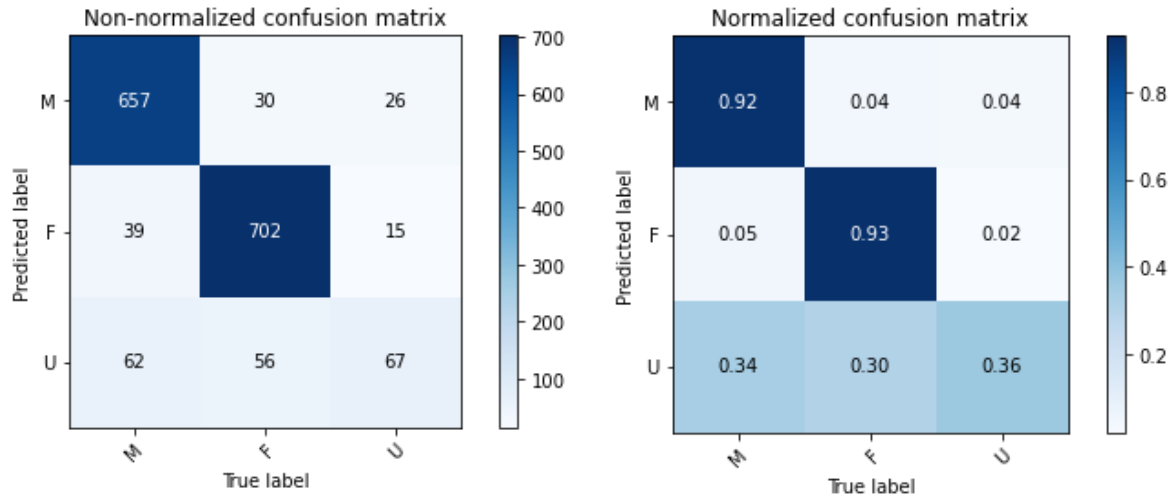


Figure 4.7: Confusion matrices for the gender classifier (SVM) non-normalized and normalized

The type of diabetes classifier achieved an accuracy of 74% with more performance measures being listed in Table 4.7. Results were similar over all classes for all measures ranging from 0.72 - 0.76. A confusion matrix is provided in Figure 4.8 indicating that a majority of misclassifications (false positives) were due to the fact that the model inferred a type of diabetes for a tweet while the true label was no type of diabetes.

	Prec	**Rec**	**F1**
Unknown	0.76	0.74	0.75
Type 1	0.74	0.72	0.73
Type 2	0.72	0.75	0.74
macro avg	0.74	0.74	0.74

Table 4.7: Performances for the type of diabetes classifier (SVM)

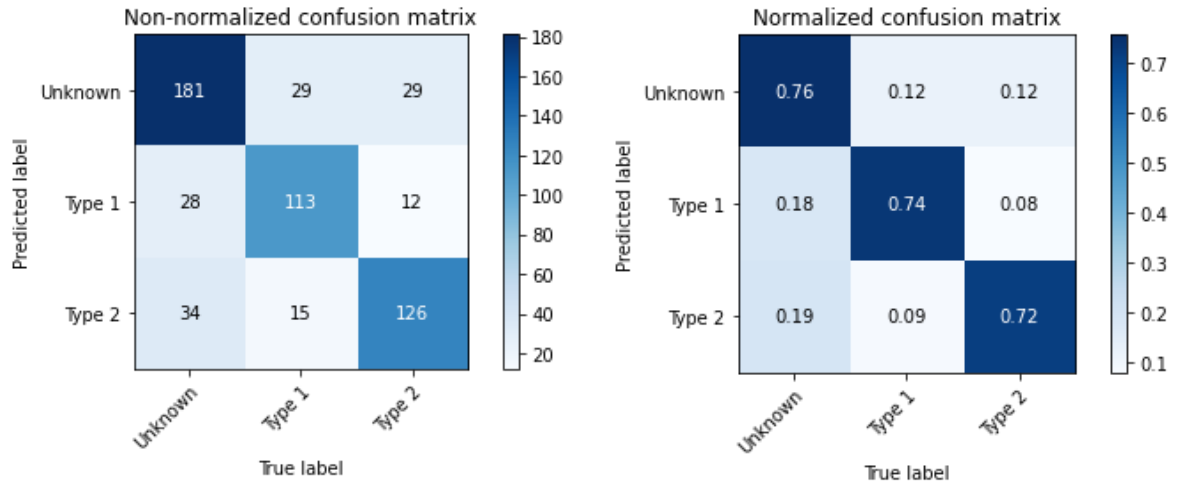


Figure 4.8: Confusion matrices for the type of diabetes classifier (SVM) non-normalized and normalized

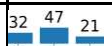
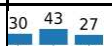
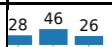
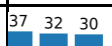
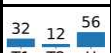
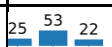
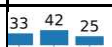
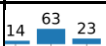
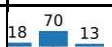
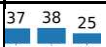
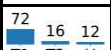
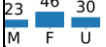
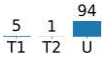
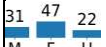
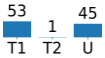
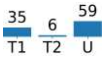
Among all diabetes-related tweets in the US, 46.407 were identified as *emotional* tweets (28% of US tweets) on which the trained gender and type of classifier were applied. This yielded 14.485 (31%) tweets written by men, 20.228 (44%) by women and 11.694 (25%) from unknown gender. In contrast, 20.285 (44%) were predicted as from people with type 1 diabetes, 4.375 (9%) from type 2 diabetes and 21.747 (47%) were from unknown type where inference was not possible.

4.3.5 Topics of interest

The heart of our study was the identification of topics of interest to characterise what people, with and talking about diabetes, share on social media. Table 4.8 outlines a detailed description of the 30 identified topics of interest obtained by the clustering. Each topic (row) is explained via a topic label that was provided by the researchers based on the most important tweets and top words. Furthermore, the most important top words, the gender and type of diabetes distribution in this topic, and a short description of the topic are provided. ANNEX 4 provides a table with sample tweets for each topic.

An over-representation of women was observed in most topics, particularly in topics 10, 23 and 26 referring to the importance of affordable insulin and the oral glucose tolerance test (OGTT). While men tend to discuss topics around diabetes-related stories (topic 6) and ‘diabetic/insulin shock’ (topic 29). Concerning the type of diabetes, discussion around ‘advocacy for affordable insulin’ (topic 10) and enjoying the exchange in the diabetes online community (topic 5) were almost

exclusively dominated by people with type 1 diabetes. The rare clusters for which people with type 2 diabetes had noticeably more tweets than on average were related to the confusion between type 1 and type 2 diabetes (topic 19) and when they tweeted about their diagnosis anniversary (topic 17). The hashtag #DSMA stands for the Diabetes Social Media Advocacy group, a community of people exchanging about diabetes-related topics, in which the price of insulin is a central concern.

No.	Topic label	Top words	Gender (%)	Diabetes Type (%)	Tweet description
1	Support/Solidarity in diabetes community	happy, birthday, day, #dsma,#t1d, insulin			- Happy birthday wishes - Affection / Support / Solidarity messages
2	Inspiring relatives living with diabetes	love, t1d, amazing, #t1d, awesome			- Inspiring friends and families when living with diabetes or helping to live a better life - Pointing to people campaigning for diabetes issues and affordable insulin
3	Sharing hope and encouraging	hope, well, #dsma, get, soon, better			- Encouraging people with diabetes - Hoping to find a cure for diabetes
4	Diabetes awareness / Support / Donation	help, love, please, us, let, awareness			- Supporting each other within the diabetes community - Raising awareness for diabetes and its complications - Donations to get insulin
5	DSMA* enjoying online support	#dsma, good, :), glad, tonight, love			- Friendly exchanges within the DSMA* online community - Being thankful for the people's effort and activity - Questions within the DSMA online community
6	Sharing diabetes related stories	good, bad, news, really, know, like			- People telling stories about diabetes or diabetes treatment - Messages about daily struggle with diabetes
7	T1D hashtags	#t1d, #diabetes, #type1diabetes, love			Mainly short, incoherent messages with several hashtags relating to type 1 diabetes, such as #t1d, #t1dlife, #t1dmom, or #Dexcom
8	Diabetes care	care, take, health, insulin, one, taking			- Talking about diabetes care or mental health with diabetes - Indignation about the healthcare system
9	Bloodsugar palette (beauty products)	#bloodsugar, got, palette, excited, love,			"Spam" about the beauty product "Blood Sugar Palette" palette
10	Advocacy for affordable insulin	#insulin4all, insulin, good, thank, love, us			- Encouraging people advocating for affordable insulin - Talking about healthcare activism
11	Life with and without diabetes	would, love, could, insulin, like, one			- Hypothetical life without diabetes - Frustration about how people without diabetes poorly understood their conditions
12	Life with type 1 diabetes	#t1d, #diabetes, love,#type1diabetes			- Messages about Dexcom devices - Seeking for encouragement - Life with type 1 diabetes
13	Glucose guardian	glucose, guardian, need, love, father			Looking for glucose guardians. This is a gender neutral term for a person who is giving money, often to buy insulin, in exchange for some favor or something in return.
14	Chatting about insulin	love, insulin, #dsma good, glucose, one			Talking about insulin
15	Insulin and insulin pump complications	insulin, pump, love, hope, got, new, feel			- Unhappiness about having to manage the insulin pump - Insulin had no effect or was bad
16	Diabetes in family	family, runs, friends, know, type, history			- Worries about getting diabetes because of family predisposition - Family support

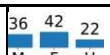
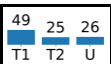
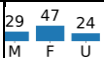
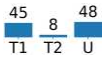
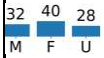
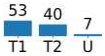
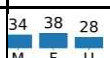
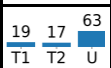
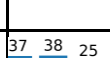
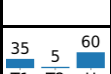
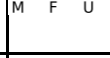
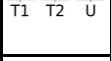
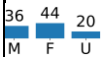
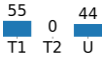
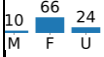
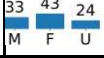
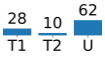
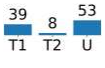
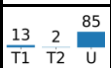
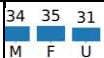
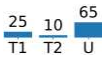
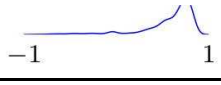
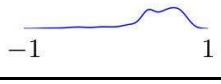
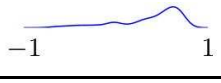
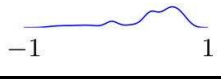
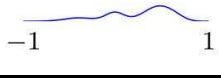
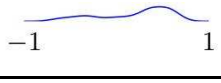
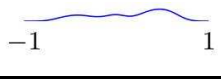
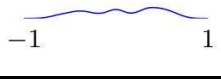
17	Diaversary	years, ago, two, type, one, insulin			Celebrating birthday of diabetes diagnosis
18	Day-to-day stories about diabetes	day, today, insulin, glucose, good, last			Day-to-day stories about diabetes or insulin adjustment
19	Confusion between T1D and T2D	type, one, two, good, love, people, feel			Frustration about people's inability to distinguish between type 1 and type 2 diabetes
20	Diagnosis	diagnosed,type, old, two, since, year, one			Talking about their own diabetes diagnosis or the one of their child
21	Glycemic instability	glucose, blood, high, sugar, insulin, levels			- Difficulty to keep blood sugar levels stable - Suffering from low or high blood sugar - Reliefs about keeping blood sugar levels normal
22	Insulin	insulin, good, like, need, people, feel			- Talking about expensive insulin and people who cannot afford treatment - Insecurity about insulin
23	Diabetes pop star Nick Jonas	nick, jonas, cried, remember, found			Talking about pop star Nick Jonas who has type 1 diabetes and wrote a song about it
24	Misunderstandings of diabetes	people, know, like, one, understand			- Complaining about the public view on diabetes - Unclear relationship for many people between food and blood sugar / diabetes
25	Frustration with insulin prices	insulin, afford, cost, cannot, insurance			- Frustration with high insulin prices - Insurance as elementary factor to get insulin
26	OGTT**	glucose, test, hour, today, tomorrow			Dreading glucose drink and glucose tests (mostly during pregnancy)
27	Parents and diabetes	mom, dad, got, died, insulin, one, bad, lost			Parents-children stories related to diabetes
28	Diabetes Distress	feel, like, hate, sick, really, bad, feeling			- Sharing feelings of sadness and depressive symptoms - Feeling scared or anxious about insulin use - Hate towards own diabetes
29	Diabetic/Insulin Shock	shock, insulin, went, going, coma, would			- Going into diabetic shock after eating sweet food - Sharing experience of going into insulin shocks
30	Diabetes-related comorbidities	anxiety, depression, pain, neuropathy, heart, cancer			Talking about diabetes-related complications such as depression, anxiety, heart diseases, nerve pain, chronic pain and other health issues

Table 4.8: Overview of the 30 topics of interest for people with or talking about diabetes and their gender and type of diabetes distributions. For each topic the following information are provided: the topic label, the most frequent words, gender (M=men, F=women, U=unknown) and type of diabetes (T1=type 1 diabetes, T2=type 2 diabetes, U=unknown) distribution and a tweet description.[79] * DSMA refers to Diabetes Social Media Advocacy online group; ** Oral glucose tolerance test

4.3.6 Primary emotions related to the topics of interest

The results related to the sentiment analysis and emotions are summarised in Table 4.9. For each topic information about the cluster size, the average sentiment analysis score (*SA score)*, the sentiment analysis distribution (SA distr.) and lastly the distribution over the primary emotions. ANNEX 5 lists the most common emotional words and emojis/emoticons per topic. The table is

ordered following the *SA score* from the most positive to the most negative topics. We have found that the most positive topics were linked to the support and solidarity within the diabetes online community (topic 1) with a *SA score* of 0.68 and 72% of emotions related to *joy* and to topic 2, "inspiring relatives living with diabetes with a *SA score* of 0.67 and 36% *joy* and 31% *love* elements. On the bottom of the table, the most negative topics were related to tweets around 'diabetes distress' (topic 28) of which the emotions *sadness* (24%), *anger* (27%) and *fear* (21%) dominated; 'diabetic/insulin shock' (topic 29) with an over-representation of *anger* (45%) and *fear* (46%) elements and concluding with 'diabetes-related comorbidities' (topic 30) with 40% *sadness,* 19% *anger* and 22% *fear* emotions expressed.

No.	Topic label	% (N)	SA score	SA distr. neg=[-1,1]=pos	Primary emotions (%) joy || love || surprise || sad || anger || fear					
1	Support/solidarity in diabetes community	1.3 (613)	0.68		0.72	0.19	0.02	0.04	0.02	0.02
2	Inspiring relatives living with diabetes	3.3 (1552)	0.67		0.36	0.31	0.21	0.05	0.03	0.05
3	Sharing hope and encouraging	1.2 (542)	0.54		0.45	0.45	0.01	0.05	0.01	0.03
4	Diabetes awareness / Support / Donation	3.2 (1499)	0.48		0.35	0.38	0.4	0.12	0.05	0.08
5	DSMA* enjoying online support	3.1 (1459)	0.47		0.48	0.26	0.06	0.09	0.04	0.07
6	Sharing diabetes related stories	3.3 (1513)	0.36		0.3	0.24	0.06	0.17	0.08	0.15
7	T1D hashtags	1.8 (812)	0.33		0.37	0.38	0.06	0.08	0.07	0.05
8	Diabetes care	2.9 (1353)	0.28		0.05	0.44	0.01	0.05	0.03	0.43
9	Bloodsugar palette (beauty products)	1.8 (818)	0.26		0.28	0.49	0.03	0.13	0.02	0.04
10	Advocacy for affordable insulin	1.1 (505)	0.22		0.24	0.29	0.09	0.15	0.13	0.1
11	Life with and without diabetes	2.7 (1239)	0.19		0.31	0.26	0.06	0.16	0.08	0.13
12	Life with type 1 diabetes	3.3 (1514)	0.18		0.3	0.27	0.04	0.16	0.11	0.11
13	Glucose guardian	0.7 (343)	0.16		0.37	0.25	0.01	0.25	0.06	0.05
14	Chatting about insulin	5.0 (2302)	0.14		0.32	0.23	0.04	0.18	0.12	0.11
15	Insulin and insulin pump complications	4.4 (2036)	0.09		0.35	0.23	0.03	0.18	0.13	0.09

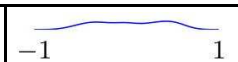
16	Diabetes in family	0.9 (436)	0.07		0.22	0.26	0.03	0.18	0.1	0.22
17	Diaversary	3.6 (1685)	0.07		0.28	0.2	0.06	0.21	0.11	0.14
18	Day-to-day stories about diabetes	5.7 (2666)	0.05		0.28	0.17	0.04	0.21	0.15	0.15
19	Confusion between T1D and T2D	3.9 (1804)	0.04		0.22	0.21	0.05	0.22	0.14	0.17
20	Diagnosis	2.2 (1023)	0.02		0.24	0.21	0.06	0.21	0.13	0.16
21	Glycemic instability	5.1 (2363)	0.01		0.22	0.19	0.04	0.24	0.15	0.17
22	Insulin	8.6 (3970)	0.01		0.27	0.18	0.05	0.2	0.14	0.16
23	Diabetes pop star Nick Jonas	0.5 (218)	-0.07		0.19	0.19	0.01	0.49	0.04	0.09
24	Misunderstandings of diabetes	7.0 (3244)	-0.08		0.16	0.17	0.04	0.29	0.14	0.2
25	Frustration with insulin prices	6.3 (2916)	-0.12		0.15	0.17	0.03	0.24	0.21	0.2
26	OGTT**	4.3 (2012)	-0.13		0.2	0.1	0.02	0.31	0.18	0.19
27	Parents and diabetes	4.0 (1859)	-0.14		0.16	0.17	0.04	0.29	0.14	0.21
28	Diabetes Distress	4.5 (2067)	-0.19		0.14	0.11	0.04	0.24	0.27	0.21
29	Diabetic/Insulin Shock	0.6 (284)	-0.33		0.03	0.02		0.03	0.45	0.46
30	Diabetes-related comorbidities	3.8 (1760)	-0.36		0.09	0.08	0.02	0.4	0.19	0.22

Table 4.9: Sentiment and emotion distributions of the 30 topics of interest of people with or talking about diabetes. For each topic the following information is provided: the topic label, the frequency of the topic, the average sentiment analysis score (SA score), the sentiment analysis distribution (SA distr.) and the frequency of primary emotions.[79]
* DSMA refers to Diabetes Social Media Advocacy online group; ** Oral glucose tolerance test

Generally, the topics using most frequently *joy* and *love* emotions turn around exchanging, supporting and showing solidarity with each other (Topics 1,2,3,4,5). The only topics revealing a higher percentage of *surprise* emotions were 'Inspiring relatives living with diabetes' (Topic 2) and 'Diabetes awareness / Support / Donation' (Topic 4). A good example of imperfect preprocessing textual data is topic 9 related to beauty products with diabetes-related names such as "Blood sugar palette". Consequently, this topic was considered irrelevant for our analyses.

4.3.7 Insulin pricing

A recurrent theme occurring in 5 of the 30 topics was Insulin pricing. In particular, topic 25 appeared as a major topic when users shared their frustration of not being able to afford insulin causing the expression of negative emotions: 24% *sadness*, 21% *anger* and 20% *fear*. Topic 13 channels this frustration when people are looking for 'glucose guardians', a gender-neutral term for a person giving money in exchange for some favor. In this context, people with diabetes typically look for glucose guardians to obtain financial support to buy life-saving insulin. Topic number 10 regroups tweets of people fighting for affordable insulin, usually represented by the hashtag '#insulin4all'. The positive emotions in this group show mutual support. The topic 22 'Insulin' and topic 4 'Diabetes awareness / Support / Donation' were broader clusters with a partial focus on users talking about expensive insulin and people who cannot afford treatment and donations as a way to finance insulin.

4.3.8 Associations between topics of interest and mean income

Regarding studying potential links between the mean household income and the topics of interest, we detected that clusters such as 'advocacy for affordable insulin' (topic 10), 'insulin' (topic 22) and 'frustration with insulin prices' (topic 25) were positively associated with the mean household city income in a way that cities with higher income were more likely to tweet about these topics ($p<0.001$ for all three topics), see Figure 4.9. Furthermore, we observed that topic 12 'life with type 1 diabetes' ($p<0.01$) and topic 21 'glycemic instability' ($p<0.05$) were positively associated with the mean income. Similarly, positive associations were identified between topic 24 'misunderstandings of diabetes' ($p<0.001$) and topic 19 'confusion between type 1 and type 2 diabetes' ($p<0.05$), both related to language use and how people exchange about type 1 and type 2 diabetes, showing that cities with higher incomes being more likely to talk about these themes.

On the other hand, negatively associated with mean city income, meaning that cities with lower income were more likely to post tweets, were detected for 'DSMA enjoying online support' (topic 5) with $p<0.001$, 'day-to-day stories about diabetes' (topic 18) with $p<0.001$ and 'oral glucose tolerance test (OGTT)' (topic 26) with $p<0.001$).

The single significant topic cities with medium mean income posted about, was 'insulin and insulin pump complications' (topic 15) with $p<0.001$.

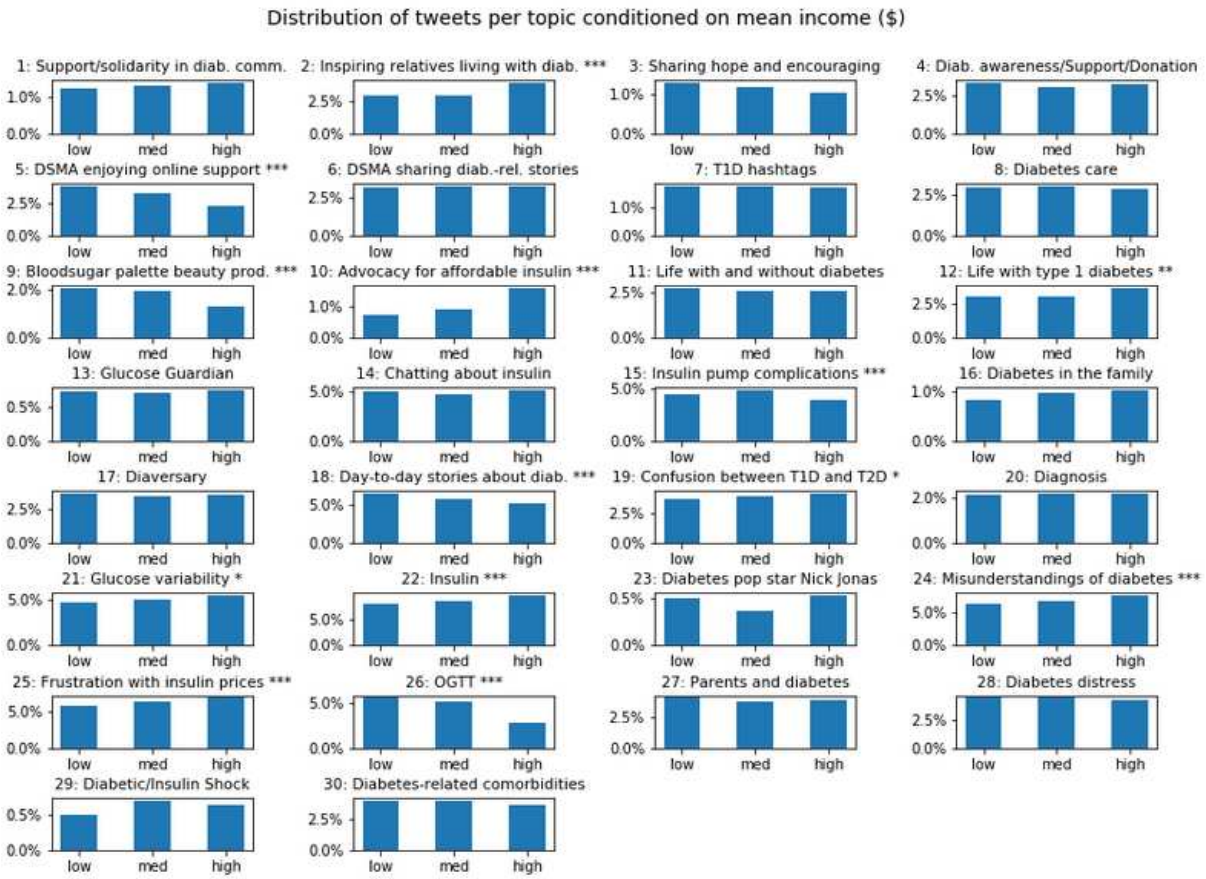


Figure 4.9: Plot of the frequency of tweets per category of mean household income at the city level (low, medium, high) for each topic. A p-value using a Chi2 test was calculated.[79]
* p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

4.4 Discussion

4.4.1 Principal findings

In this work we showed that Twitter is a useful tool to capture and describe key diabetes-related topics and emotions related to those clusters. Based on a strict preprocessing pipeline we solely focused on relevant tweets, namely tweets with personal and emotional content which are no jokes and coming from the US. Our main findings suggest that on one hand among the diabetes online community there exists a lot of mutual support and solidarity with multiple tweets containing *joy*

and *love* elements. On the other hand, users shared *fear*, *anger* and *sadness* emotions when it comes to insulin pricing and diabetes-related complications and comorbidities. Also, significant frustration was caused by people's inability to distinguish between the different types of diabetes. Moreover, we found that cities with higher incomes were more likely to discuss topics around affordable insulin, insulin pricing, the language use related to diabetes or glycemic instability. Whereas discussions in cities with lower mean income were more driven around daily stories related to diabetes, exchanging in the online community (DSMA) and the oral glucose tolerance test.

4.4.2 Comparison with the literature

To the best of our knowledge, this was the first study capturing information concerning key diabetes-related concerns based on social media data. For this reason, we were not able to confront and compare the results with others. Nevertheless, a close study in terms of methodology published by Vydiswaran et al. studied relationships between food-related discussions on Twitter and neighborhood characteristics.[252] Similarly to our study they divided tweets into personal and non-personal, using an external tool, worked on geo-located tweets and associated a count-based emotional score to each tweet. However, their study had a regional focus of tweets from the Detroit Metropolitan area by only gathering already geotagged tweets and thus avoiding an external geolocation step as in our case. To identify themes, they manually coded a small subset of 1537 tweets whereas we used unsupervised machine learning to be able to work on a larger dataset.

Further, we demonstrated gradients between topics related to diabetes and the household income level of the city, similar to Nguyen et al. who illustrated that people living in zip codes with high percentages of happy and physically active tweets had lower obesity prevalence based on geolocated Twitter data.[253]

Other studies documented the importance of understanding emotions and self-control (regulation of thoughts, emotions and behavior) for health outcomes in people with diabetes based on self-reports.[254] Coccaro et al. confirmed our results that diabetes distress is linked with negative emotion.[255] Hagger et al. examined diabetes distress among adolescents with type 1 diabetes and observed a significant proportion experienced elevated diabetes distress which is often associated with suboptimal glycemic control.[256] Another study suggested that positive emotions such as hope and curiosity may play a protective role in the development of a disease.[257] Furthermore, it has been shown that higher levels of emotional distress are related with poor self-care in type 2 diabetes.[140] Iturralde et al. showed that anxiety is highly comorbid with depression among individuals with type

2 diabetes.[258] Those studies emphasize the relation between emotions, diabetes and diabetes distress and thus align with our results demonstrating that people with or talking about diabetes frequently share emotions, diabetes- and diabetes distress related topics and concern on social media.

4.4.3 Insulin pricing

A major concern among tweets in the USA was insulin pricing affecting 5 out of 30 topics with the whole spectrum of emotions present. The positive emotions (*joy, love)* occurred in tweets referring to solidarity in the fight for affordable insulin within the diabetes community. Those tweets were often accompanied by the hashtag '#insulin4all', a campaign that unites the diabetes community around the accessibility to treatment for everyone.[259] Negative emotions (*sadness, anger, fear*) were present when people shared their frustration regarding insulin prices, the access to insulin and identifying sources of insulin including *glucose guardians* or donations, which encompass major obstacles for people with diabetes.[76,77] Blanchette et al. also stressed the fact that greater financial stress and psychological factors have detrimental impacts on self-management outcomes during emerging adulthood in people with type 1 diabetes and suggest advocating for policy changes to support improved self-management outcomes.[260]

We have found associations of topics focusing on insulin pricing to be more frequent in cities with high mean incomes. One can not directly conclude that this indicates that people living in cities with a high mean household feel more concerned about insulin prices, but rather that they probably have a greater ability to tweet around this issue. Numerous tweets geolocated in high mean household income cities included the hashtag '#insulin4all'. Key challenges for a global and fair access to insulin are known.[261]

Some reasons for the high cost of insulin are the presence of a vulnerable population depending on insulin to survive; the quasi monopoly of three companies dominating the market and the immense lobbying power of these companies.[262] To our knowledge, we are the first to demonstrate and quantify the extent of the crisis in the USA, with respect to insulin pricing, on a large sample of people with or talking about diabetes based on social media data. Besides, we highlighted different emotions and fears associated with the crisis around insulin pricing.

4.4.4 Strengths and limitations

Various strengths pervade this study. The first and certainly one of the most crucial advantages of using social media data is the fact that information is expressed spontaneously and in real-time. This can be described as an open digital space with a flat role hierarchy for information sharing, and online communities development. As a consequence, potential biases occurring in traditional and observational studies such as the responder bias could be minimized because there is no hierarchy between the parties. Secondly, the large number of people and the wide variability in their profiles were analyzed. Third, we developed an innovative methodology to focus on relevant (personal, emotional, non-joke) geolocated tweets from the USA to identify topics of interest and emotions shared within topics. And lastly the approach is able to capture trends in the diabetes online community and socioeconomic factors which can be associated with social media at the ecological level.

At the same time, several limitations need to be mentioned. First, users expressing diabetes-related concerns on Twitter may not be representative of all people with diabetes. Nonetheless, it has been suggested that this can be partially offset by the large number of people sharing data in the first place, a major strength of digital epidemiology.[27] Despite the observation of large variability in the profiles, we spotted an over-representation of people with type 1 diabetes and women when compared to known diabetes epidemiology literature when in reality the vast majority of diabetes cases are type 2 (90%).[1] A potential explanation of the greater percentages of type 1 diabetes cases might be the younger demographics of Twitter users.[263] An alternative hypothesis might be that type 1 diabetes involves more care, more devices and more challenging medication resulting in more frustration which is shared on Twitter compared to type 2 diabetes. Nevertheless, our results should be interpreted in the context of a social media population solely. A second limitation is the fact that the performances, in particular the precision, of our classifiers was not perfectly precise, meaning that there is no guarantee that all our tweets were actually posted from people with diabetes sharing personal content and frequently it was not possible to determine the gender or type of diabetes. Third, we did not account for clinical and environmental factors which might have helped tease out these factors. For each topic a label was provided which is not exclusive. By zooming into the clusters it could have been possible to explore more subtopics and sub-themes. This could be a potential future investigation. The fourth limitation is the bias introduced in our geolocation by inferring a location based on user-provided locations which might not be their true locations. Fifth, dealing with the advent of sarcasm and irony in emotion detection is still an open research question.

Finally, causal inference between the mean household income per city and the topics of interest of people living in the geolocated city cannot be made as it is subject to ecological fallacy.

A natural extension of this work is to extend our analyses to other countries and other languages.

4.4.5 Conclusion

In this study, we explored diabetes-related topics and their associated emotions. A central concern we found was insulin pricing and it comes with negative emotions such as *sadness, anger* and *fear.* We demonstrated the feasibility of capturing emotions, concerns and interests of individuals in real life and showed that it is an efficient way of augmenting psychosocial, behavioral and epidemiological research. Moreover, we found that frequent diabetes distress related topics are discussed on social media which were not included in current diabetes distress scales such as the frustration related to insulin access; emotions and fear regarding diabetes and its complications or being annoyed about the general confusion between type 1 and type 2 diabetes. With this work we hope to encourage future studies to consider social media data as complementary online information.

Social media is a direct source to capture information about people with diabetes feelings, emotions, beliefs, worries, and fears linked to diabetes, diabetes treatment and complications among the large and active diabetes online community. Thus, it provides a useful observatory for diabetes issues. The utilization of Twitter analysis on diabetes could inform the public debate about diabetes issues and so help to contribute directly to public and clinical decision making. Social media data will help develop policies and interventions that consider key concerns among people with diabetes to ultimately improve health outcomes.

CHAPTER V: CAUSALITY DETECTION

This chapter addresses the second thesis objective of identifying cause and effect relationships in diabetes-related tweets.

The findings of this chapter have been submitted for peer-review to an international journal.

5.1 Introduction

The accessibility of a large volume of social media data offers researchers a new perspective in studying causal relations in patient-reported outcomes expressed in real-time. Specifically, causes of health problems, concerns and emotions can be detected as well as effects of risk factors and actions. Extracting cause-effect relations has gained popularity in biomedical knowledge discovery[264,265] or emergency management.[266] In the particular case of Twitter, causal relations have been extracted for adverse drug reactions,[267] characterising causes for stress and relaxation tweets[268] or for causal association extraction related to insomnia and headache.[80]

The aim of this work was to extract spans of text as two distinct events from individual diabetes- and diabetes distress related tweets under consideration, such that one event directly or indirectly impacts another event. We categorized these events as cause-event and effect-event depending upon the expressed context of each tweet.

The majority of causal relation extraction approaches examine *explicit* causality in text,[80,265,269] when cause and effect are explicitly stated by a causal link (e.g. so, hence, because of, since), causative verbs (e.g. break, kill), conditional (e.g. if.. then..) or causative adverbs and adjectives.[270,271] An example for an *explicit* cause-effect pair is “diabetes causes hypoglycemia” with the word “cause” as explicit causal link. Whereas *implicit* causality is more complicated to detect such as in “Cannot sleep #insomnia, #overthinking” with cause “#overthinking” and the effect “Cannot sleep”. Asghar divided the cause-effect relationship detection into two types of methods: pattern/rule-based or machine learning-based methods.[272] While rule-based methods require the manual construction of a set of linguistic and syntactic rules,[80,265,273] machine learning systems can use a small subset of patterns to then train an algorithm to automatically find these language patterns.[274–276] An

advantage of machine learning based approaches is their capability of exploring implicit relations and to generalise beyond the seen examples.[270]
More advanced deep learning models have also been applied to extract causal relations.[277,278] Ponti and Korhonen improved the implicit causality identification using a feedforward neural network.[279] Another interesting approach, leveraging the transfer learning paradigm, combining both explicit and implicit cause-effect extraction is provided by Khetan et al.[280] They fine-tuned pre-trained transformer based *BERT* language models[64,65] to detect "cause-effect" relationships using publicly available datasets such as the adverse drug effect dataset.[281]

In a similar spirit, the objective of the present work is to identify both explicit and implicit multi-word cause-effect relations on diabetes and diabetes-distress related tweets.
In this work we focus on intra-sentential causality where cause and effect lie in a single sentence and can be a span of text or a single word.

5.2 Material and methods

An overview over the workflow is visualised in Figure 5.1. The first step consisted in preprocessing the diabetes-related tweets to remove noise and concentrate solely on tweets with relevant content. The second step focused on the analysis, starting with the detection of tweets containing causal information (e.g. observation, opinion, concerns, etc.), also referred to as *causal tweets* resp. *causal sentences*. Causes and corresponding effects were then extracted from *causal tweets*. Finally, cause-effect pairs were clustered, described and visualised.
These steps will be elaborated throughout the following sub-sections.

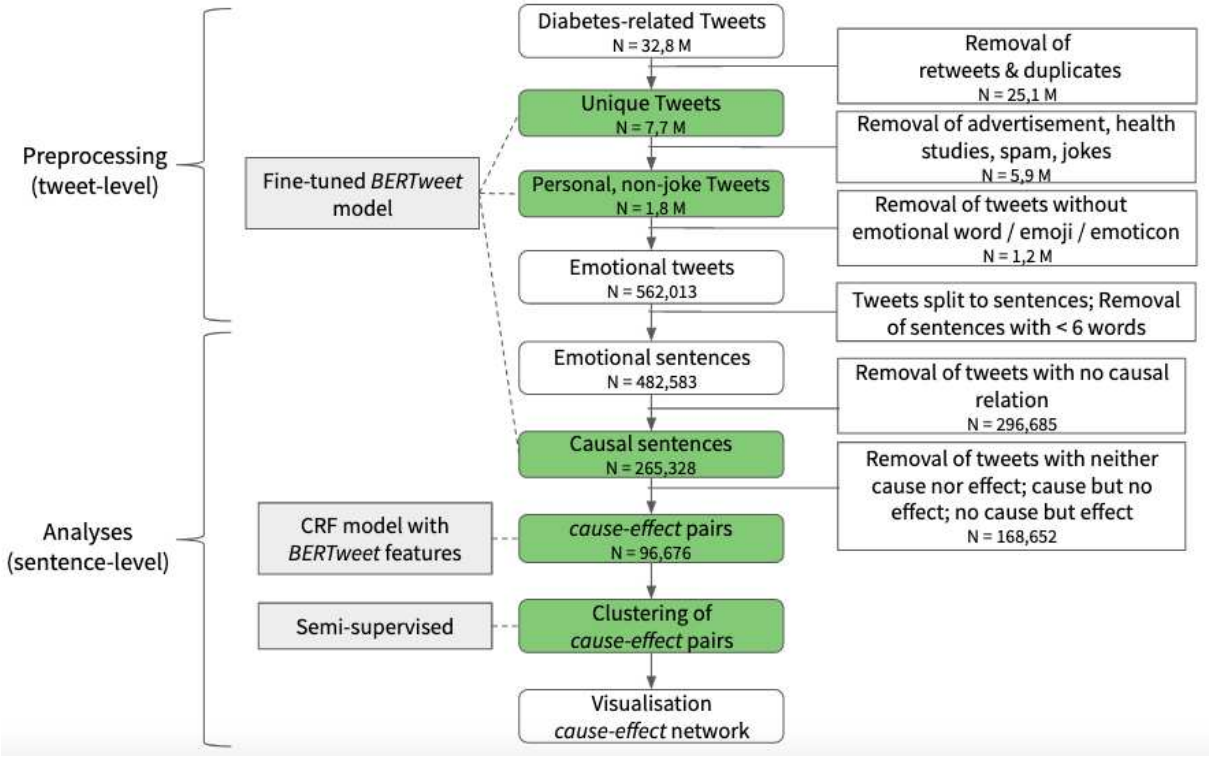


Figure 5.1: Workflow over the cause-effect extraction. The steps in green color include machine learning methods.[282]

5.2.1 Data collection

The foundation for this work built 32 million diabetes-related tweets in English collected between April 2017 and January 2021. For more details on the data collection, refer to chapter 3.2.1.

5.2.2 Data preprocessing

The data preprocessing was heavily influenced by the preprocessing pipeline introduced in chapter IV. In a first step, retweets and duplicates were removed leading to a database of 7.7 million unique tweets. Secondly, tweets with *personal* content were identified, with feelings, emotions and opinions shared by people with or talking about diabetes. In consequence, *institutional* tweets communicating advertisement, news or health information were excluded. For this purpose we leveraged the transfer learning paradigm and fine-tuned an already pretrained transformer[65] based language model *BERTweet.*[78] *BERTweet* was trained on 850 million english tweets (16 billion word tokens ~80GB) collected from January 2012 to August 2019 following the *RoBERTa* pre-training procedure.[188]

To fine-tune the binary classification task, a linear layer was added on top of the last *BERTweet* layer utilizing the *transformer* package of Huggingface.[283] The same training data as for the *personal tweets* classifier in previous work (Chapter IV) was harnessed and extended to a total of 4,303 tweets (1539 personal, 2764 institutional). The dataset extension helped to counteract potential concept drifts, which describe underperformances of machine learning models when trained on data in the past and applied to contemporary data.[96]
For the model training, the same preprocessing steps were undertaken as those used for the pre-training of the language model, including translating emotion icons into text strings and converting user mentions and url links into special tokens. The performance to identify tweets with personal content achieved 91,2% accuracy, 86,2% precision, 90,9% recall and F1 score of 88,5%. Applying this classifier on all unique tweets led to 2.5 million tweets with personal content.

Analogously to chapter IV, jokes around diabetes, a common phenomenon, were considered out of scope for this study. In a similar spirit, *BERTweet* was fine-tuned for joke detection. For this purpose, the joke dataset from chapter IV, was as well augmented to a total of 1,648 tweets (486 jokes, 1,162 non-jokes) to address the concept drift. Performance in detecting jokes reached an accuracy of 90,4%, precision of 78,5%, recall of 90,8% and F1 of 84,2%. The joke detection further reduced the dataset to 1.8 million personal, non-joke tweets.

A last preprocessing step emphasized our study focus on diabetes distress and consequent psychological factors and emotions. Identical to chapter IV, solely tweets containing an emotional element (emojis/emoticon, or emotional word) were captured for this intent, resulting in 562,013 tweets containing personal, non-joke and emotional content.

5.2.3 Data annotation

With the completion of the preprocessing cycle, we can now turn to the analysis part. The identification of both causal tweets and cause-effect pairs relied on a causal dataset we manually labeled. We did not restrict ourselves to a specific area of diabetes-related causal relationships and include potentially all types. For this purpose 5,000 randomly chosen tweets, from the 562,013 preprocessed tweets, were manually annotated with the four columns to be labeled: *Intent* describing the intent of the tweet such as if it is a question (“q”), has a negation in either cause or effect (“neg”), has multiple causes (“mC”) or effects (“mE”), is a joke/irony/sarcasm (“joke”), has

multiple sentences (“mS”) or has multiple sentences but cause-effect pair is in a single sentence (“msS”); the second and third column describes the cause respective effect, if existing, where several causes/effects may occur. The last column *Causal association* is a binary variable specifying if a cause-effect pair exists.

Tweet	Intent	Cause	Effect	C.A.*	Explanation
Diabetes causes me to have mood swings		Diabetes	mood swings	1	**Possible causal association**
I just want to eat . I hate #diabetes	msS	#diabetes	hate	1	**Possible causal association** related to diabetes distress
Scary, have a diabetic daughter but I read thousands of people a year die in UK just from flu so why panic over corona .	mS			0	**Non-diabetes or diabetes distress related relationship.** “flu” is non-diabetic related
I'm back ! Had two strokes and recover now. Have high blood pressure and diabetes. :-)	mS			0	**Unclear cause-effect relationship** Not clear if “High blood pressure” or “diabetes” caused the stroke
Not sure if I am up since 3:30 to watch Titanic or because of my anxiety over my glucose test is what keeps me up 😬		glucose test	anxiety	1	**Chaining cause-effect relationship A->B->C** event A: glucose test event B: anxiety event C: been up since 3:30 => label the relationship which is closest to our study objective: diabetes and diabetes distress
My 14 year old daughter is Type 1 = malfunctioning pancreas , meaning not enough insulin being made to regulate 😟	msS; mE	Type 1	malfunctioning pancreas; not enough insulin	1	**Negation** Negation in a cause/effect is considered being part of the cause/effect as it does not alter the meaning
it is not true to think that insulin makes you feel so bad 🙃	neg	insulin	feel so bad	0	**Negation** Negation is not part of cause/effect and alters the meaning

Table 5.1: Sample tweets in different label scenarios. The tweets are fictive to ensure privacy but represent similar real tweets. *C.A.: Causal association

Table 5.1 illustrates some example tweets (adapted from the original tweets to ensure privacy). For a more detailed explanation on the annotation please refer to our annotation guidelines in ANNEX 6.

Labeling cause-effect pairs is a complex task. In order to assess the reliability of the labeling, two researchers labeled 500 tweets independently and calculated Cohen’s kappa score, a statistical

measure expressing the level of agreement between two annotators.[284] A score of 0.83 was attained being interpreted as an *almost perfect* agreement according to Altman, and Landis and Koch.[285,286] Disagreements were discussed between both authors and one author carried on labelling the remaining 4,500 tweets, resulting in 5,000 labeled tweets.

5.2.4 Models

Cause-effect associations were identified in a two step process. A first model was trained to predict if a sentence contains a potential cause-effect relation (causal sentence); a second model extracted the cause and associated effect from the causal sentences. The first model can be interpreted as a barrier filtering non-causal sentences out. Non-causal sentences may have either a cause, an effect, none of them, but not both. To simplify model training, we hypothesized that cause-effect associations only exist in the same sentence and all sentences with less than 6 words were removed due to a lack of context. In consequence, both models operate on a sentence and not tweet level. Further difficulties in this environment were that *causes* and *effects* could be multi-word entities.

5.2.4.1 Causal sentence detection

Detecting causal sentences is a binary classification task. We built a feedforward network on top of a pre-trained *BERTweet* language model consisting of two fully connected linear layers (FCLL) with dropout layers (probability: 0.3) and completed by a softmax layer which translates the model predictions into probabilities, see Figure 5.2.

The initial training data consisted of 5,000 imbalanced tweets and, after splitting tweets into sentences, led to 7,218 non-causal sentences and 1,017 causal sentences. To balance the data imbalance, the categorical cross entropy loss function was parametrized by class weights to penalise mispredictions for causal sentences stronger. Training parameters were: adam optimizer with *epsilon* of 1e-8; scheduler with linearly decreasing learning rate and 0 warm up steps; learning rate of 1e-3 and was trained for 35 epochs with early stopping. Training data was stratified and separated as 90% training and 10% test data. 20% from the training set were extracted for validation. Batch size for training and validation was 16, and for the test set 32.

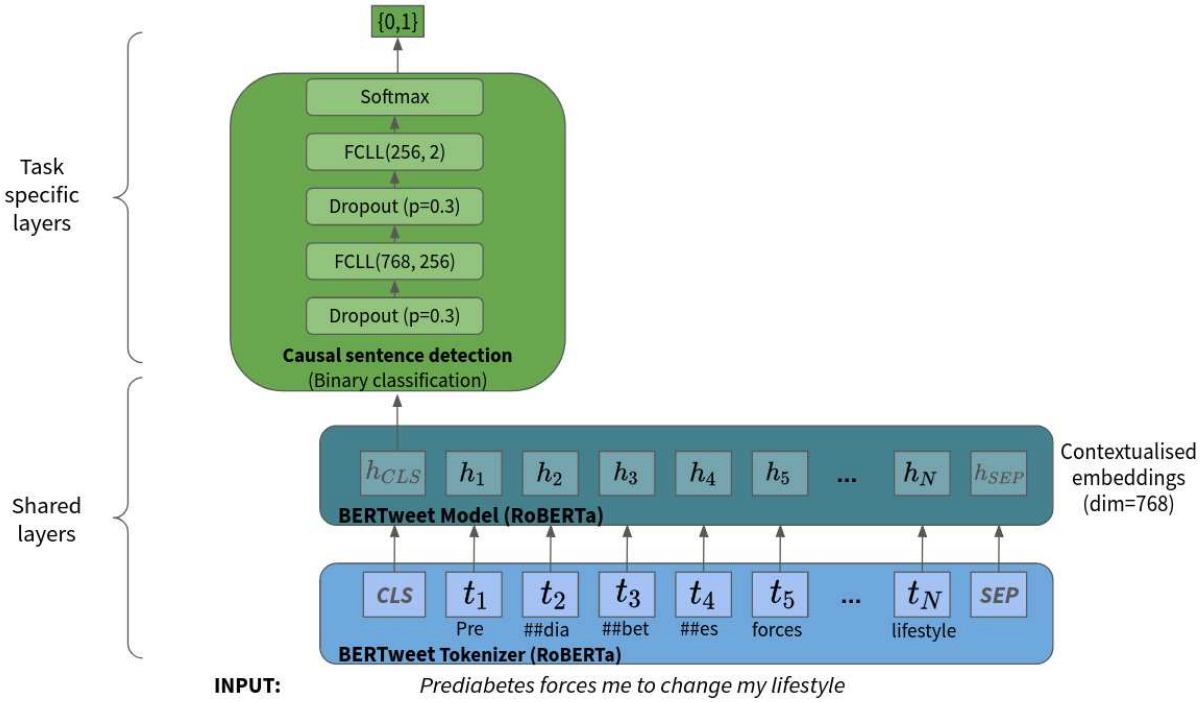


Figure 5.2: Model architecture - Causal sentence detection.[282]

Data augmentation through active learning

On the one hand the data imbalance and on the other hand the few positive training examples per cause-effect pair, as causes and effects could possibly be linked to any diabetes-related concept, motivated us to adopt an active learning approach to increase the training data volume.

Figure 5.3 exemplifies our active learning routine which increases the training data iteratively.

During the first iteration, the causal sentence classifier starts by training on the initial training set of 5,000 tweets (which are split into sentences for training). The second step consists in applying the trained classifier on 2,000 randomly chosen unlabeled tweets, which led to a set of positive predictions (causal sentences) and negative predictions (non-causal sentences). In a third step, only the causal sentences were manually examined to correct possible misclassifications. The non-causal sentences set remained unaltered, and in consequence potential misclassifications remained in the non-causal sentences which were then considered noisy. In a last step, both the corrected causal sentence set and the uncorrected non-causal sentence set were combined and added as new training data to the already labeled database, resulting in an updated training database of 7,000 tweets. The whole procedure was repeated four times and enabled us to augment the labelled data much faster and efficiently than without active learning as it allowed us to focus on the few positive samples.

The final resulting training set served then to train the causal sentence classifier as well as the cause-effect extraction model.

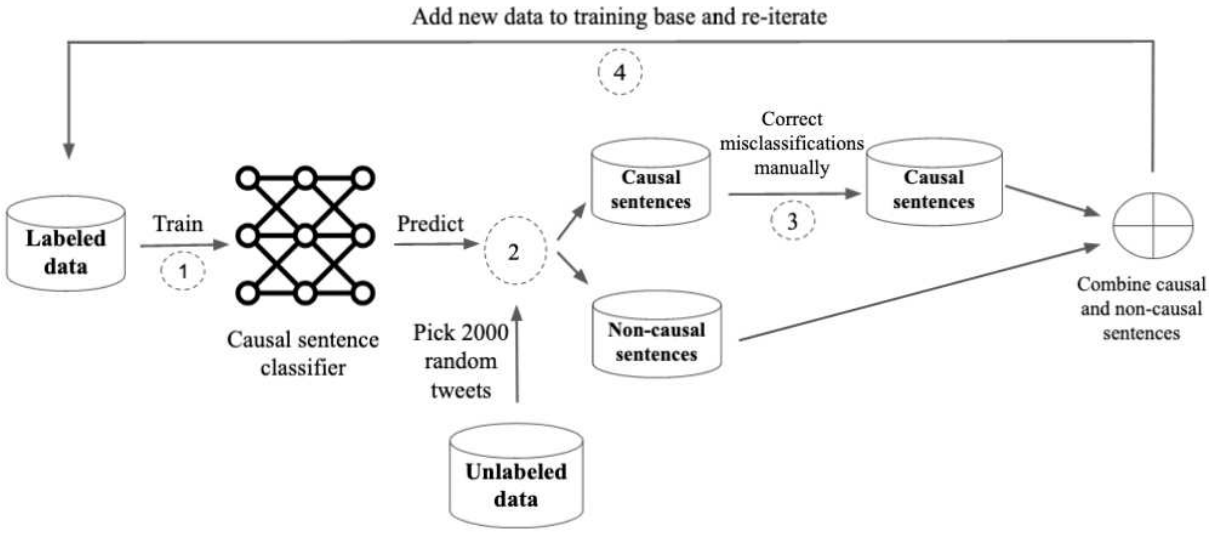


Figure 5.3: Active learning loop to augment the training set in a time-efficient fashion.[282]

5.2.4.2 Cause-effect associations

The identification of cause-effect associations was casted as an event extraction, or named entity recognition task, i.e. assigning a label *cause* or *effect* to a sequence of words. We encoded the manually labeled *causes* and *effects* in a IO tagging scheme based on the common tagging scheme BIO (Beginning, Inside, Outside), introduced by Ramshaw and Marcus.[201] In this context, "B-C" defines the beginning of a cause, analogous "B-E" defines the beginning of an effect. "I-C" denotes inside the cause and "I-E" inside the effect. The outside tag "O" symbolized that the word is neither cause nor effect. We hypothesized that the attention mechanism in the *BERTweet* model delivers sufficient information about the positions of each word allowing us to remove the beginning tags "B-C" and "B-E". This resulted in a simplified tagging scheme and learning task for the model with only three possible classes. Instead of applying BIO tagging, IO tagging can yield to superior results in named entity recognition tasks.[287] A sample IO tagging scheme is illustrated below with cause *Prediabetes* and effect *change my lifestyle*:

Sentence :	*Prediabetes*	*forces*	*me*	*to*	*change*	*my*	*lifestyle*
IO tags :	*I-C*	*O*	*O*	*O*	*I-E*	*I-E*	*I-E*

An additional difficulty in this setting is that a word can be both cause or effect depending on the context. In the sentence "*Prediabetes forces me to change my lifestyle"* the word "*Prediabetes"* acts as *cause* whereas in "*Limited exercising may lead to prediabetes"* it takes the role of the *effect*. A further complication originated from the fact that the task was considered open domain, as *causes* and *effects* were not restricted to one specific topic, but could be linked to any concept in our target domain: diabetes. Those reasons made it challenging to create a representative training set, as most *cause-effect* pairs occurred rarely.
This complexity drove us to test several model architectures, compare Figure 5.4 for an overview:

1. **BERT_FFL:** Pre-trained *BERTweet* language model and on top two feed forward layers with a dropout of 0.3 before, followed by a softmax layer. For the model training the cross entropy loss function is selected and weighted by the class weights to penalise mispredictions for causes and effects stronger.

2. **WE_BERT_CRF:** Single CRF layer with *BERTweet* embeddings as features augmented by discrete features such as if the word is lowercase, digit or the word length. The CRF function is implemented with the python package *sklearn-crfsuite*[288] based on *CRFsuite.*[289] As parameters for the CRF function, the default algorithm "Gradient descent using the L-BFGS method" was chosen and coefficients for L1 and L2 regularization were 0.1.

3. **FastText_CRF**: Similarly to WE_BERT_CRF, with the difference that *BERTweet* embeddings were replaced by *FastText* embeddings in the feature vector for each word. The same *FastText* embeddings from chapter IV were applied. They were trained on our own diabetes-related tweets and in such well adapted to our use case

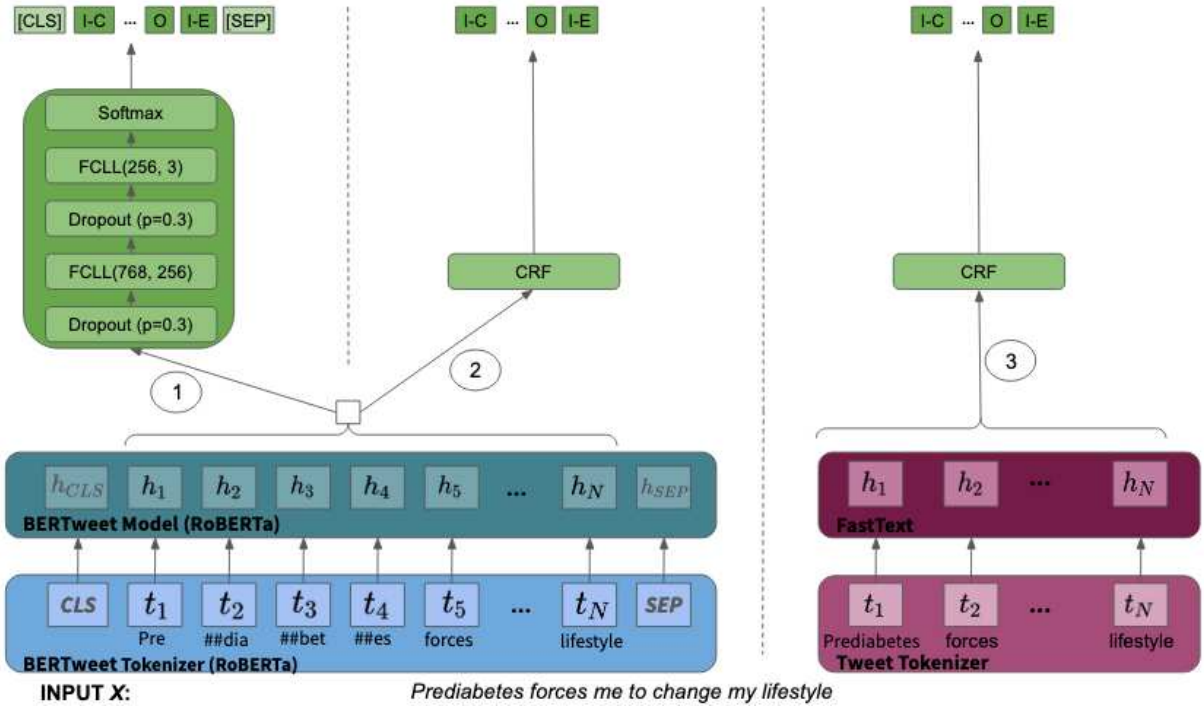


Figure 5.4: Model architectures - Cause-effect identification.[282]

5.2.5 Clustering of causes and effects

To allow effective network analyses and visualisation, *causes* and *effects* were regrouped into similar concepts (clusters). A semi-supervised and time-efficient approach was adopted in which 1,000 *causes* and 1,000 *effects* were randomly selected and two researchers manually grouped these into clusters such as “diabetes”, “death”, “family”, “fear”, as well as summarized clusters into “Parent clusters” to simplify understanding. In an automatic procedure, the remaining *causes* and *effects* were then compared to each element of all clusters, using cosine similarity, and associated to the cluster containing the most similar element. A similarity threshold of 0.55 was determined experimentally; if a cause/effect achieved a cosine similarity smaller than this threshold for all elements, a new cluster was created for this cause / effect. This procedure resulted in 1,751 clusters of causes and effects. To remove noisy clusters through potential misclassifications, exclusively clusters with a minimal number of 10 cause/effect occurrences were kept, leading to 763 clusters. Note, the order of documents might affect the results, as different clusters might have been created. ANNEX 7 provides an overview over the 100 most frequent clusters, where automatically added clusters had the value “Other” as “Parent cluster”.

Finally, these clusters were visualised in an interactive cause-effect network developed in D3 to obtain a deeper understanding about the cause-effect relationships.

5.2.6 Software

Python (version 3.8.8) and the deep learning framework PyTorch (version 1.8.1) were used to implement above-mentioned methods. The algorithms are open-sourced under following address: https://github.com/WDDS/Causal-associations-diabetes-twitter/

5.3 Results

Results were obtained on the basis of 482,583 sentences which were extracted from splitting the 562,013 personal, non-joke and emotional tweets into sentences and excluding questions and sentences with less than 6 words.

5.3.1 Model performance

5.3.1.1 Causal sentences

The performances to detect causal sentences for the imbalanced dataset are shown in Table 5.2. Each active learning round was trained on more data.

Round	N° sent. train	N° sent. test	Accuracy	Precision	Recall
0	6,024	837	64.5	58.0	67.4
1	7,536	1,047	67.7	61.2	71.6
2	8,804	1,223	67.7	60.3	66.3
3	10,284	1,429	65.4	60.0	68.8
4	11,861	1,648	71.0	61.0	67.8

Table 5.2: Performance measures (macro) for each round of more training data

Highest accuracy was reached in round 4 with 71%, whereas the precision of round 1 and round 4 were almost equal with 61.2% respectively 61.0%. The model of round 4 was applied on all remaining tweets to detect causal sentences, as it was trained on the largest training data set, including difficult causal examples missed by earlier models, and was thus better at identifying

complex causal sentences. The active learning strategy enabled us to increase the training data much quicker than without active learning and without loss in performance. This resulted in a clean database of 265,328 causal sentences with most noisy sentences removed.

5.3.1.2 Cause and effect detection

The active learning strategy allowed us to extend the dataset to 2,118 causal sentences, i.e. containing both cause and effect. This dataset was split into 90% train and 10% test set; and 20% of the training set served as validation set. The performance of the different cause-effect models is listed in Table 5.3.

The best performing model was the CRF model with *BERT* embedding features (WE_BERT_CRF) achieving a precision, recall and F1 of 0.68. Remarkably, WE_BERT_CRF outperformed a fine-tuned *BERT* model, which is considered the gold standard of current NER tasks. A potential explanation for that is that *BERT*-based models make local decisions at every point of the sequence taking the neighboring words into account before its decision. In a situation like ours, with strong uncertainty on all elements, due to the complexity of the task, a single CRF layer model, leveraging *BERT* features, makes global decisions using the local context of each word and thus maximizes the probability of the whole sequence of decision better.
Besides, the CRF model with simple *FastText* embeddings achieved strong results, probably due to the fact that word embeddings were specifically trained on this diabetes corpus.

In consequence, the WE_BERT_CRF model was applied on all causal sentences leading to a dataset of 96,676 sentences with *cause* and associated *effect* predicted.

Models		Prec	Rec	F1
BERT_FFL	I-C	0.48	0.46	0.47
	I-E	0.20	0.48	0.29
	O	0.91	0.77	0.83
	macro	0.53	0.57	0.53
WE_BERT_CRF	I-C	0.63	0.61	0.62
	I-E	0.49	0.49	0.49
	O	0.93	0.93	0.93
	macro	0.68	0.68	**0.68**
FastText_CRF	I-C	0.59	0.57	0.58
	I-E	0.45	0.38	0.41
	O	0.92	0.94	0.93
	macro	0.65	0.63	0.64

Table 5.3: Performance measures for each of the four architectures

5.3.2 Cause-effect description

The largest clusters are summarized on the left side in Table 5.4 and on the right side the most frequent cause-effect associations are listed, excluding the largest cluster “Diabetes” as it will be studied separately. The cluster “Diabetes” is the largest one with 66,775 occurrences of “Diabetes” as either cause or effect (ex.: #diabetes, diabetes, diabetes mellitus) followed by “Death” with 16,989 (ex.: passed away, killed, died, suicide, etc.) and “Insulin” (ex.: insulin, insulin hormone, etc.) with 14,148. From the 30 largest clusters, 6 refer to nutrition, 4 to diabetes and 3 clusters to each of insulin, emotions and the healthcare system.

Most frequent clusters			*Most frequent cause-effect-associations (excluding cluster "diabetes")*		
Parent cluster	**cluster**	**N**	**cause**	**effect**	**N**
Diabetes	diabetes	66,775	unable to afford insulin	death	1,246
Death	death	16,989	insulin	death	1,156
Insulin	insulin	14,148	type 1 diabetes	fear	1,054
Diabetes	type 1 diabetes	11,693	type 1 diabetes	death	999
Emotions	fear	10,160	rationing insulin	death	805
Glycemic variability	hypoglycemia	9,547	type 1 diabetes	insulin	751
Symptoms	sick	6,549	OGTT*	sick	584
Nutrition	overweight	5,186	type 1 diabetes	hypoglycemia	578
Diabetes	type 2 diabetes	4,909	insulin	hypo	545
Complications & comorbidities	neuropathy	4,481	insulin	fear	534
Healthcare system	medication	4,389	type 1 diabetes	insulin pump	436
Diabetes Technology	insulin pump	4,307	finance	death	423
Nutrition	nutrition	4,230	type 1 diabetes	sick	400
Emotions	anger	4,149	insulin	sick	385
Health	OGTT*	4,053	insulin	finance	367
Blood pressure	hypertension	3,782	type 1 diabetes	anger	356
Healthcare system	finance	3,767	insulin	medication	305
Nutrition	reduce weight	3,589	insulin	anger	296
Insulin	unable to afford insulin	3,381	OGTT*	fear	293
Nutrition	diet	3,325	type 2 diabetes	death	293
Emotions	sadness	3,153	type 2 diabetes	fear	290
Glycemic variability	hyperglycemia	3,144	hypertension	death	286
Diabetes	suffer	3,132	overweight	death	280
Diabetes Distress	depression	2,810	type 1 diabetes	finance	277
Healthcare system	hospital	2,721	hypoglycemia	insulin	272
Diabetes Distress	stress	2,681	hypoglycemia	sick	263
Nutrition	sugar	2,369	affordable insulin	death	262
Nutrition	fasting	2,363	insulin	insulin pump	255
Insulin	rationing insulin	2,244	complications	death	248
Health	gestational diabetes	2,076	insulin	sadness	240

Table 5.4: Most frequent cause-effect clusters (left) and associations (right) with the number of occurrences. The cause-effect associations on the right side exclude the cluster "Diabetes".
*OGTT: Oral glucose tolerance test

The largest cluster "Diabetes" mainly occurs as a cause and its most frequent effects ("Death", "fear", "sick") are visualised in Figure 5.5. From the 30 most numerous effects for "Diabetes", 6 were

related to “Nutrition” and 5 to “Complications & comorbidities” and 3 to each of “Diabetes distress”, “Emotions” and “Healthcare system”.

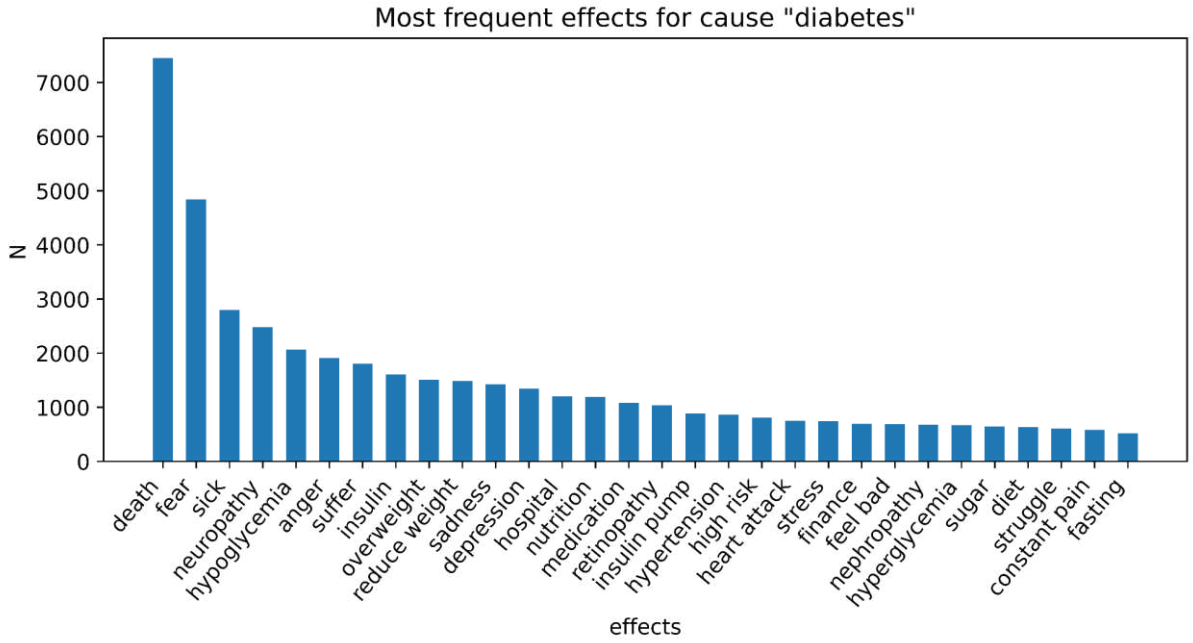


Figure 5.5: Most frequent effects for the largest cluster “Diabetes”.[282]

The interactive visualisation in D3, with filter options, was published under the following link: https://observablehq.com/@adahne/cause-and-effect-associations-in-diabetes-related-tweets.
We invite the interested reader to play with the graph and explore the different cause-effect associations. A sample graph of this visualisation is illustrated in Figure 5.6 showing only cause-effect relationships with at least 250 occurrences to ensure readability. It is eye-catching that “death” appeared to play such a central role as *effect* with various causes hitting at it: “unable to afford insulin”, “rationing insulin”, “finance”, “insulin”, “Type 1 diabetes (T1D)”. Other central clusters were “Type 1 diabetes” acting as cause for “insulin pump”, “insulin”, “hypoglycemia (hypo)”, “sickness”, “finance” and emotions “anger” and “fear”, where latest has the strongest association; or the node “Insulin” mostly relating as cause to “sickness”, “medication”, “finance”, “death”, or “hypoglycemia” and “fear” and “anger”.

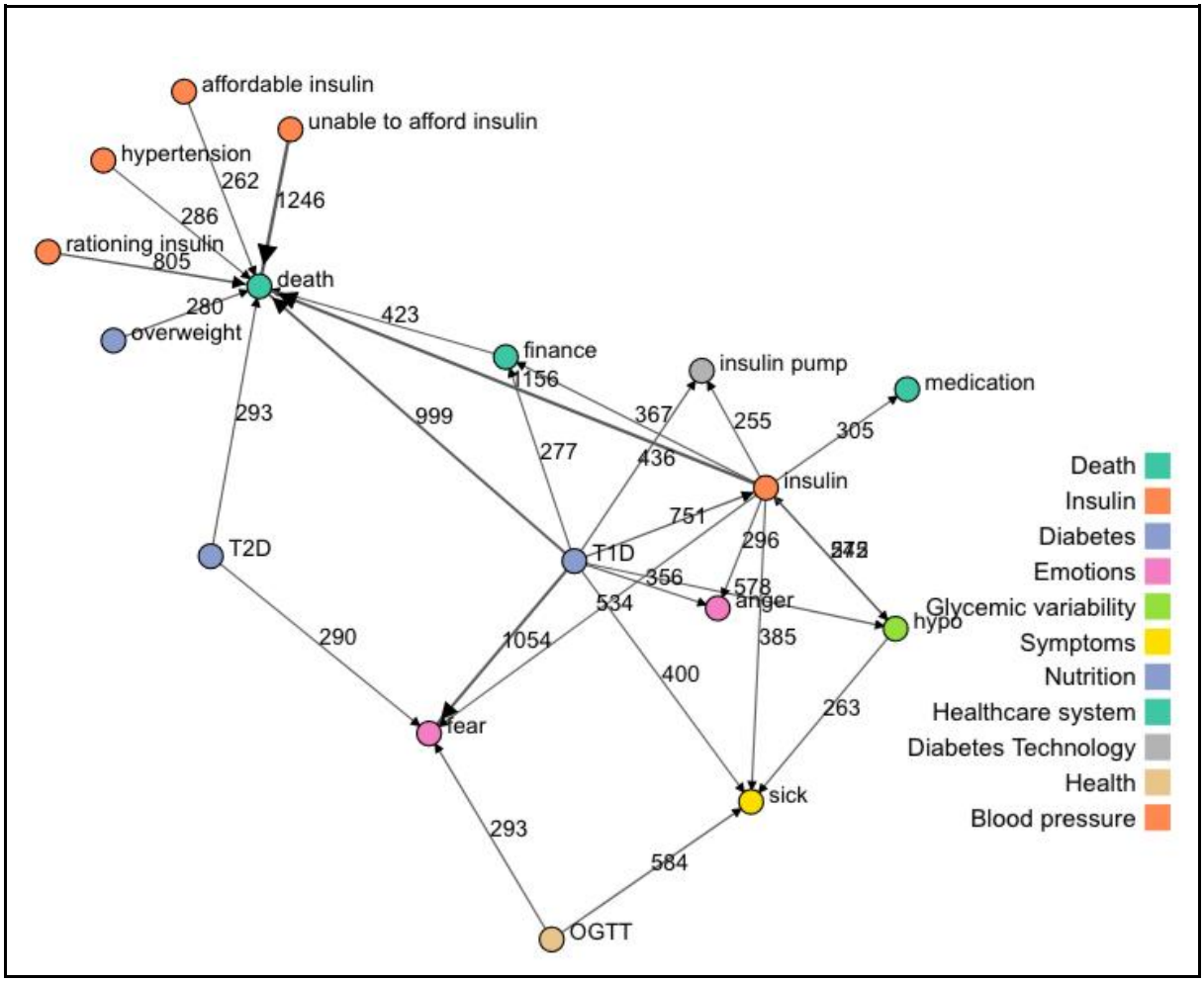


Figure 5.6: Cause-effect network with a minimum number of associations (edges) of 250.[282] Accessible under: https://observablehq.com/@adahne/cause-and-effect-associations-in-diabetes-related-tweets

5.4 Discussion

5.4.1 Principal results

Our findings demonstrate the feasibility of extracting both explicit and implicit cause and associated effects from diabetes-related Twitter data. We have shown that by adopting the transfer learning paradigm and fine-tuning pre-trained language models causal sentences could be detected. Furthermore, we demonstrated that simply fine-tuning a *BERT*-based model does not always outperform more traditional methods, but a simple conditional random field augmented by *BERT* features led to a more effective detection of cause-effect relationships. Given the challenging task and imbalanced dataset, the performance measures were satisfying. But, we would like to

emphasize the importance of data quality and the availability of a large training dataset which represent major points to address in future investigations to improve the performance. Semi-supervised clustering enabled us to create a cause-effect network which can be explored by the interested reader in an interactive visualisation. “Diabetes” was identified as the largest cluster acting mainly as the cause for “Death” and “fear”. Besides “Diabetes”, a central cluster was detected in “Death” acting as an effect for various causes related to insulin pricing, a link already detected in earlier works.[79] In addition, “Type 1 diabetes (T1D)” and “Insulin” were frequently mentioned.

5.4.2 Comparison with the literature

Former studies have already examined causality on Twitter data. Doan et al. focused on the three health-related concepts: “stress”, “insomnia”, “headache” which were considered the effects and identified causes using manually crafted patterns and rules.[80] Yet, they only studied explicit causality and excluded causes and effects encoded in hashtags and synonymous expressions.[80] We, on the contrary, tackled both explicit and implicit causality and included causes and effects encoded in hashtags and exploited synonymous expressions through the use of word embeddings. Kayesh et al. proposed a novel technique based on neural networks which uses common sense background knowledge to enhance the feature set, but they similarly concentrated on the simplified version of explicit causality in tweets.[269] Bollegala et al. developed a causality-sensitive approach for detecting adverse drug reactions from social media using lexical patterns and in consequence targeted explicit causal patterns.[290] Dasgupta et al. proposed one of the few deep learning approaches, due to the unavailability of appropriate training data, which leverages a recursive neural network architecture to detect cause-effect relations from text, but also only targeted explicit causality.[291] An approach tackling both explicit and implicit causality is provided by Khetan et al. who used a *BERT*-based model on an already existing labeled corpora not based on social media data, contrary to our approach.[280] Recently they further extended their work of explicit and implicit causality understanding in single and multiple sentences in clinical notes.[292]

To the best of our knowledge, this is the first study targeting both explicit and implicit cause-effect relationships on diabetes-related Twitter data.

5.4.3 Strengths and limitations

The present work demonstrates various strengths. First, leveraging powerful language models enabled us to detect a large number of tweets containing *cause-effect* relationships having led to the identification of 20% (96,676 / 482,583) of tweets with cause-effect associations, contrary to other approaches which were able to identify causality in less than 2% of tweets.[80] Second, contrary to most previous work, we tackled both explicit and implicit *causal relationships*, an additional explanation for the higher number of *cause-effect* associations compared to other studies focusing only on explicit associations.[80] Third, we could avoid defining manually crafted, imperfect patterns to detect causal relationships by relying fully on automatic machine learning algorithms. Fourth, operating on real-time social media data that is expressed spontaneously offers the opportunity to extend our knowledge from an alternative data source which might complement traditional epidemiological data sources.

A strong limitation is that *cause-effect* relations are expressed in tweets and this cannot be used for causal inference as the Twitter data source is uncertain and the information shared can be opinion or observation. A future point of investigation could be the testing of our results in a more traditional setting, such as cohort-questionnaires. Another shortcoming is that the performance of our causal algorithms to detect *cause-effect* pairs is not perfect. However, the overall process and the vast amount of data minimizes this issue. The lack of recall is counterbalanced by the sheer amount of data and the lack of precision is counterbalanced by the clustering approach in which non-frequent causes or effects are discarded.[81] Manual labeling of causes and effects in a dataset is a highly complicated task and we would like to emphasize that mislabelings in the dataset may occur. Enhancing data quality surely is a primary point to address to further improve performance. The causal association structures learnt by the model from the training set, might not generalise entirely when applied on the large amount of Twitter data. Moreover, during the active learning strategy only positive samples were manually corrected and negative samples were included to the training base without verification which certainly added additional noise to the model and could be improved in future investigations. Besides, we would like to highlight that the diabetes related information shared on Twitter, may not be representative for all people with diabetes. For instance we observed a bigger cluster of causes/effect related to type 1 diabetes compared to type 2 diabetes, which is contrary to the real world.[1] A potential explanation for that is the age distribution of Twitter

users.[293] But due to the large number of tweets analyzed, a significant variability in the tweets could be observed.

5.4.4 Conclusion

In this study, an innovative methodology to identify possible cause-effect relationships among diabetes-related tweets was developed. This task was challenging due to several aspects: addressing both explicit and implicit causality; multi-word entities; the fact that a word could be both cause or effect; the open domain of causes and effects; the biases occurred during labeling of causality; and the relatively small dataset for this complex task. These challenges were overcome by augmenting the dataset via an active learning loop and the exploitation of modern deep learning architectures. The feasibility of our approach was demonstrated using *BERT*-based architectures in the preprocessing and causal sentence detection. A combination of *BERT* features and CRF layer were leveraged to extract causes and effects in diabetes-related tweets which were then clustered in a semi-supervised approach. The visualisation of the cause-effect network based on Twitter data can deepen our understanding of diabetes, in a way of directly capturing patient-reported outcomes from a causal perspective.

CHAPTER VI: CLINICAL DECISION SUPPORT SYSTEM

This chapter details the third objective in which the literature summarization part in the clinical decision making process was tackled. An innovative methodology has been developed to efficiently structure, analyze and visualise scientific literature with a focus on interpretability, ease-of-use for health practitioners without programming skills and the capability of handling large data corpora.

The content of this paper has been peer-reviewed and accepted for publication in the international journal: Journal of Medical Internet Research.[294]

6.1 Introduction

6.1.1 Clinical decision support systems for literature summary

Efficient literature search skills are necessary for healthcare professionals to properly exercise evidence-based medicine (EBM) for clinical decision making.[295] Nonetheless, limited available time, knowledge and skills hinder the application of EBM, explaining why only one in every five medical decisions is based strictly on evidence.[296,297] A way out of this dilemma lies in clinical decision support systems (CDSS) powered by AI solutions. CDSS assist health professionals in providing targeted knowledge, patient information and other health information.[298] But significant challenges for an efficient CDSS persist such as the questions of how to use clinical knowledge, for instance extracted free text information; how to transform the clinical knowledge into a usable form and how to mine large clinical databases.[299] In short, there is a need for high-quality decision support capacities for health professionals to efficiently interpret the exponentially growing data,[299–301] including electronic health records, laboratory results, doctor-patient interactions, social media or biomedical literature[79,302–304] to enhance clinical knowledge in the decision process.

6.1.2 Machine learning to analyse textual data

Machine learning and natural language processing methods translate these health data into actionable knowledge such as disease phenotype identification,[305] hospital readmission[306] and decision support.[307] Despite those advances, major obstacles preventing machine learning from being more widely deployed in the healthcare domain are their lack of accessibility to non-technical users and uncertainties about their reliability on real-world data.[308,309] Particularly challenging for non-technical users is setting up modern machine learning systems which require specialized hardware and complicated software dependencies.[310] Besides, lack of interpretability and explainability hamper the adoption in real practice.[311,312] Another limitation is the lack of effort of integrating available expert knowledge into existing machine learning models to improve interpretability.[313]

Typical approaches to analyze unstructured textual information are *topic models* such as Latent Dirichlet Allocation (LDA),[160] connecting documents that share similar patterns, or leveraging word embeddings (*Word2Vec*, *FastText*, *BERT*) in combination with a clustering algorithm (e.g. K-means).[79] Yet, they suffer diverse shortcomings. Most of these methods require the definition of the desired number of topics prior to the model training.[314] Topic models lack scalability, and are memory-intensive on large text corpora.[315] Moreover, the synthetic nature of topics, created by topic models, do not explicitly correspond to prior knowledge of humans concerning topics in the corpus domain.[315] In addition, these models are static systems in the sense that it is not possible to add new documents to the model without a complete retraining. A last limitation is that the closed form of these algorithms prevents a user from interacting and influencing the topic exploration.

6.1.3 Objectives

The third thesis objective was to propose an online decision support algorithm for non-experts, people without computer or data science knowledge, to discover topics of interest and classify unstructured health text data.

Specifically, the contributions were: 1) Proposition of a single methodology for biomedical document classification and topic discovery which improves interpretability; 2) Creation of an open-source tool for users without programming skills, able to run on machines with limited calculation

power and on big data clusters; 3) Evaluation of this methodology on a real world use case to show its capability to reach near state-of-the-art performance while addressing aforementioned limitations.

Ultimately, this methodology shall analyze a broad set of different biomedical texts in various scenarios. The evolution of scientific interest over time, based on publication, or of public health opinion in social media can be studied due to the dynamic nature of our approach allowing the easy injection of new documents to the model. Free text on surveys or cohort participants' opinions in free-text content such as questionnaires can be investigated. Or simply the classification of biomedical documents such as medical records, reports or patient's feedback.

The aim of this study was not to set a new benchmark regarding performance but instead tackle existing limitations in NLP approaches, usability and interpretability of CDSS in the healthcare domain to ultimately enhance the literature exploration in the clinical decision making process.

6.2 Material and methods

6.2.1 Data

Preliminary tests on short text messages from Twitter showed satisfying results. The scope of this work centered around the evaluation of longer text messages consisting of several phrases or paragraphs. Freely accessible biomedical text corpuses with existing hierarchical labels are scarce. One of the few sources are the scientific abstracts from Pubmed, which therefore served as test corpus for this methodology. As described in the Data section of the Chapter III, Pubmed abstracts were downloaded from the Courtesy of the U.S. National Library of Medicine which provides them with an already classified hierarchical structure, namely Medical Subject Headings (MeSH) codes.[56,316]

Diabetes MeSH code hierarchy	MeSH code	n°
Diabetes Mellitus	*C19.246*	-
Diabetes complications	C19.246.099	5000
Diabetes Angiopathies	C19.246.099.500	3026
Diabetes Foot	C19.246.099.500.191	4424
Diabetes Retinopathy	C19.246.099.500.382	5000
Diabetes Cardiomyopathies	C19.246.099.625	386
Diabetes Coma	C19.246.099.750	97
Hyperglycemic Hyperosmolar Nonketotic Coma	C19.246.099.750.490	97
Diabetes Ketoacidosis	C19.246.099.812	1308
Diabetes Nephropathies	C19.246.099.875	5000
Diabetes Neuropathies	C19.246.099.937	3662
Diabetes Foot	C19.246.099.937.250	4424
Fetal Macrosomia	C19.246.099.968	1282
Diabetes Gestational	C19.246.200	5000
Diabetes Mellitus, Experimental	C19.246.240	5000
Diabetes Mellitus, Type 1	C19.246.267	5000
Wolfram Syndrome	C19.246.267.960	228
Diabetes Mellitus, Type 2	C19.246.300	5000
Diabetes Mellitus, Lipoatrophic	C19.246.300.500	85
Donohue Syndrome	C19.246.537	39
Latent Autoimmune Diabetes in Adults	C19.246.656	16
Prediabetic State	C19.246.774	1261

Table 6.1: Diabetes MeSH codes with the number of documents for each MeSH code. [294]

Within the downloaded abstracts 293,889 contained a diabetes related MeSH code classified under the concept of Diabetes Mellitus as defined in Table 6.1. A memory limitation of 30GB Ram obliged

us to further reduce the dataset for our analyses to be able to compare our approach with competing algorithms. For this reason, only abstracts with a single MeSH code were kept and a threshold of maximal 5,000 abstracts per MeSH code was fixed. Several tests have been performed to determine this threshold experimentally. These tests started with a threshold of 1,000 abstracts per MeSH code and each next test this threshold increased by 1,000 until a maximal memory capacity was reached for a limitation of 5,000 abstracts per MeSH code. The final number of abstracts for our analyses was 50,911, published from as early as 1949 up to 2020. Table 6.1 details the number of abstracts for each MeSH code.

The underlying word embeddings for this work were trained on biomedical texts from MEDLINE/PubMed in the context of the BioASQ challenge and thus were well adapted to our situation.[317]

6.2.2 Methods

In the following we propose a novel methodology to interactively cluster documents in a hierarchical tree form following a *top-down* style. In an iterative process, a user alters and manipulates this tree until a desired user-defined clustering solution of documents is found via an interactive user interface. A global overview of this procedure is outlined in Figure 6.1.

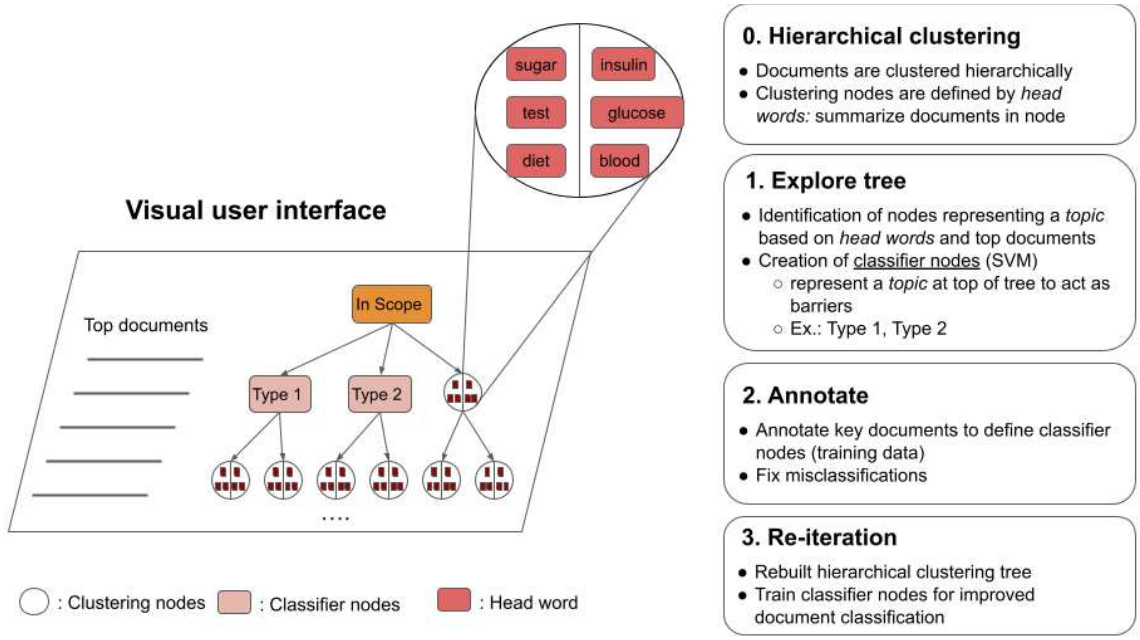


Figure 6.1: Overview of user interaction via the visual user interface.[294]

Documents start entering the tree at the root node *In Scope* and are then processed individually to construct the tree from the top to the bottom. A crucial design choice was that documents were streamed one-by-one, avoiding the storage of all documents in memory, or even the need to know their total number which leads to a radical gain in memory consumption. A practical consequence is that this design choice adds a dynamic dimension to the model, enabling us to study cluster dynamics and evolutions as new documents can continuously be added. The tree consists of two types of nodes, *classification nodes* and *clustering nodes.* A *classification node* is a binary machine learning classifier representing a theme or concept created by the user and defined by positive and negative annotations which serve as training data for the classifier. *Classifier nodes* are situated at the top of the tree and act as a barrier letting only documents pass to the underlying nodes if they match the defined concept (e.g.: “Type 1” and “Type 2” *classifier nodes* in Figure 6.1). The role of a *clustering node* is to split (cluster) documents on the basis of automatically identified *head words* which best describe the documents having passed this node (e.g. *head words* “sugar”, “test”, “diet” and “insulin”, “glucose”, “blood” for the zoomed node in Figure 6.1).

In short, *classifier nodes* orient a document from the root node to a certain tree branch where the document continues to be clustered in *clustering nodes*.

The initial tree is *binary* which splits each node into two children. Through user interaction it is possible to create several child nodes, which will be detailed in the next sections. Besides, the tree

is not equilibrated in the sense that some nodes stop splitting earlier than others and thus leaf nodes at different depths of the tree exist.

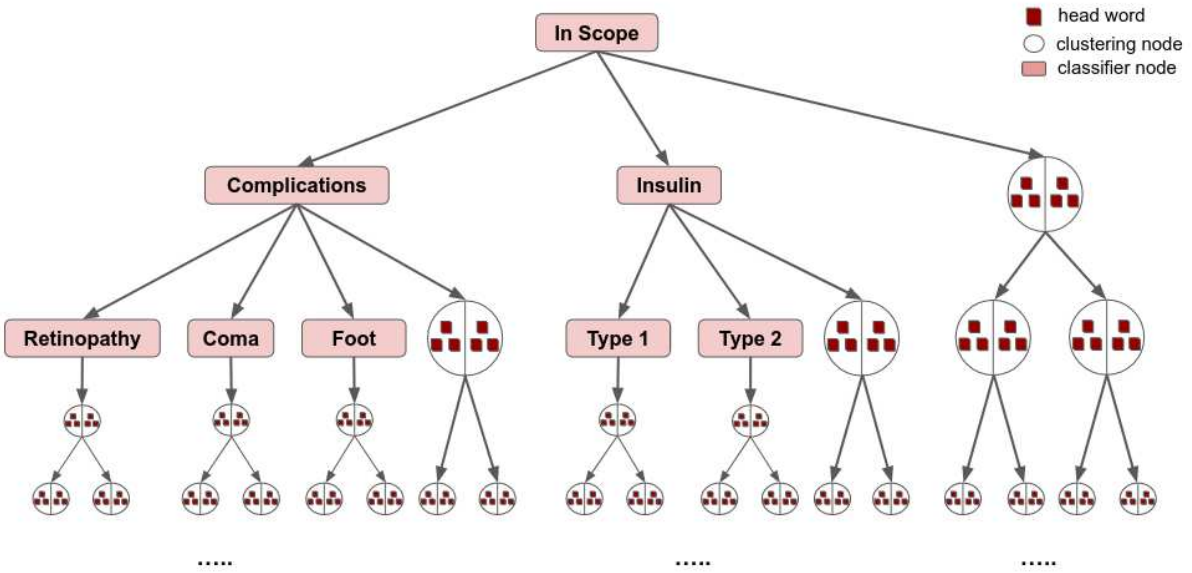


Figure 6.2: Tree structure with *classification nodes* and *clustering nodes* after several iterations.[294]

Figure 6.2 highlights a potential sample tree; a user aims to build after several iterations with *classifier nodes* at the top of the tree to filter documents in corresponding branches in which *clustering nodes* continue to cluster documents to be explored by the user.

The root node in each tree, also referred to as "*In Scope*", is a token-level *classifier node* and represents a first barrier for irrelevant documents based on the user's interests. Here, the idea is to incorporate potential domain knowledge of the user into the system, by defining words the user judges relevant and important for his use case, so called positive annotations. For instance, if a user aims to discover diabetes-related topics, potential positive annotations might be *diabetes, insulin* or *hypoglycemia*. A built-in list of stopwords (e.g. *and, of, or, for,* etc.) serves as negative annotations, as it is hypothesized that those words contain only little discriminative information, but can be altered by the user as well. These positive and negative words serve as training data for the *In Scope* classifier and thus to filter irrelevant documents out. In experimental tests we observed that only a few annotations are required to define a performing *In Scope* classifier.

At step 0 the documents enter “In Scope” and are then processed one-by-one to build the tree in a fully-automatic procedure which will be detailed in the following sections, compare Figure 6.1 and iteration 0 in Figure 6.3. The initial constructed tree is composed of one *classifier node*, “In Scope” and all underlying nodes are *clustering nodes*. Once the initial tree is built, the user starts exploring the tree (Step 1) and tries to detect *clustering nodes* that potentially summarize a specific topic or concept using the interface. The interface provides information about the *head words* and most important documents for each *clustering node* that guide the user in his choice. Let us assume a user has identified a *clustering node,* as a node regrouping documents referring to the theme of “Type 2 diabetes”, he then labels this node with a corresponding name (Ex.: “Type 2”) which provokes the creation of a *classifier node* at the top of the tree with this name via the user interface. In step 2 the user picks sample documents which refer to the theme (“Type 2”), the positive annotations and some sample documents which do not relate to the theme, the negative annotations for the *classifier node*. The selected annotations serve as training data for the underlying machine learning classifier of the *classifier node* to predict if a document refers to the corresponding theme. In step 3, the end of an iteration, all *classifier nodes* are trained and a new hierarchical tree is constructed by streaming all documents again (iteration 1 in Figure 6.3). The intuition behind the newly created and trained *classifier nodes* is that they directly group together documents corresponding to the user-defined concept and in this way act as a filter. Documents having passed such *classifier nodes* continue to be clustered enabling the user to identify possible sub-themes. In the following iteration cycles, a user carries on in exploring the tree, identifying more themes resulting in more *classifier nodes*, choosing more training documents and fixing potential misclassifications to improve the classification performance.

In each iteration several *classifier nodes* can be created. *Classifier nodes* are always children of another *classifier node* at the top of the tree and each has a single *clustering node* child where clustering of documents continues, compare Figure 6.2.

With this semi-guided, active interaction between the user on the one hand and the system on the other hand, the performance of the *classifier nodes* is improved in each iteration, leading to a smarter regrouping of similar documents and ultimately resulting in a model that converges towards a desired user-defined clustering solution.

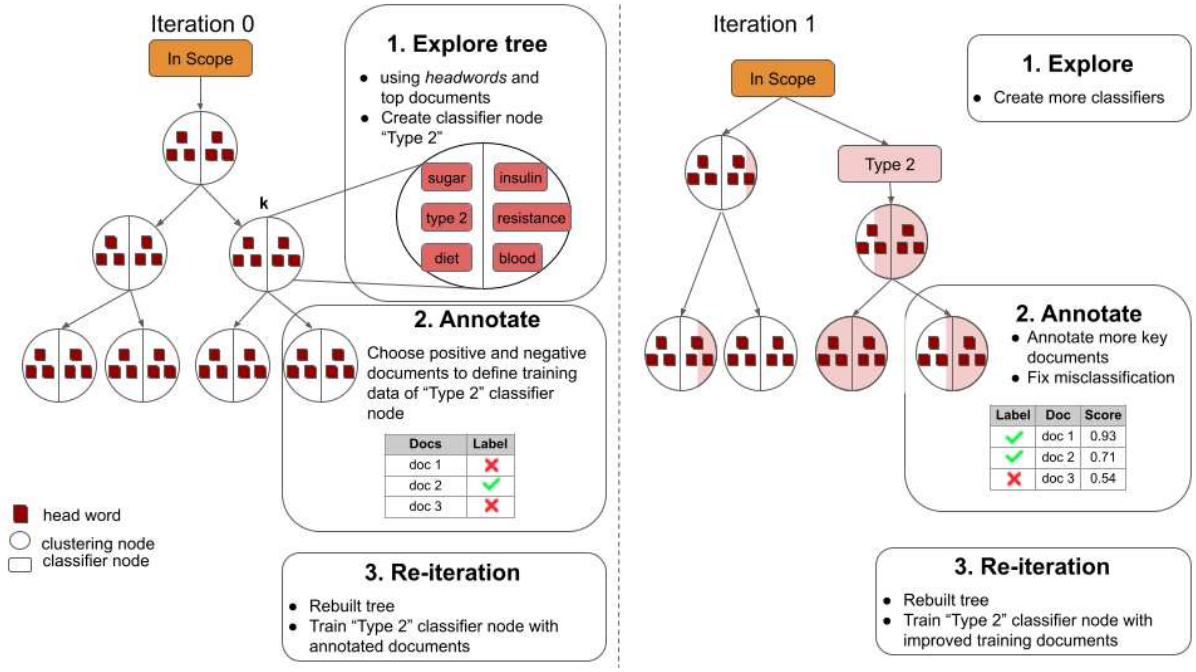


Figure 6.3: Iterative user interaction via the user interface repeating the three steps of exploring, annotating and re-iterating. For simplification purposes in iteration 1 no more *classifier nodes* are created. Generally, in a real-case scenario several classifiers are defined in the first iterations. In iteration 1, the red fill in the nodes indicates the percentage of documents in those node referring to the theme "Type 2".

The outcome of this interactive procedure is: 1) a calibrated visualization framework for a given text corpus; 2) a cascade of classifiers directing new documents to the most appropriate tree branch.

Word embeddings again built the foundation of this methodology, by transforming tokens and *head words* into a vector format.

The next paragraphs outline the approach in four sections:

1) A novel hierarchical clustering algorithm processes documents in a streaming nature
2) User-defined *classifiers* targets concept and topics
3) The visual user interface builds a connection between user and system to explore the tree, annotate documents and correct misclassifications
4) A fully parallelizable interactive and iterative process results in an accelerated convergence and minimized user annotation effort using the interpretable tree structure in combination with Active learning

The whole methodology is implemented in the programming language Scala[215] using the large-scale data processing framework Apache Spark.[216] The visual interface was developed with JavaScript and the visualisation library D3.[214]

6.2.2.1 Hierarchical clustering

The idea behind the *head words* of a *clustering node*, which best describe the documents having descended this *clustering node*, is that a user can read the *head words* and obtain an immediate understanding of the underneath documents. This substantially improves interpretability, as the user can easily retrace and understand the clustering path of documents by reading the *head words* via the interface. In this context, a well-chosen set of *head words* for a given *clustering node,* means more formally that the sum of the *head word's* word embeddings must be as close as possible to the sum of word embeddings of all tokens having passed through the node.

From a high-level point of view when a document arrives at a *clustering node,* it is compared to both *clustering node children,* to determine the path the document takes, and associated to the child it scored highest for. This score is calculated by aggregating scores of the document tokens that have arrived at this node, to its closest *head word* in the child nodes. During this process it is also tested if replacing a token by its most similar *head word,* in the node the document got associated to, yields an improvement. This *head word* replacement is essential, as the iterative clustering process continues until the optimal set of *head words* is found for each *clustering node*.

To illustrate this in formulas we define: the default number of *head words* N_h is 6, where N_h / 2 *head words* are associated to child 1 and N_h / 2 to child 2; p_j^l are the *head words* with $j \in \{0, .., N_h\}$ being the index of the *headword*; and $l \in \{1,2\}$ specifies the child of this *head word*.
To determine which path a document takes at a given *clustering node*, following three steps are evaluated:

1. <u>*Token scores*</u>: For all tokens t_i in the document return the highest cosine similarity to a *head word* of each child 1 and child 2:

$$score_{ij}^{1} = max_j\ sim(t_i, p_j^1)$$
$$score_{ij}^{2} = max_j\ sim(t_i, p_j^2)$$

2. <u>*Child score*</u>: For both children calculate the aggregated token scores over the *head words*

$$childScore_1 = avg_{j \in \omega_1}\ avg_{i \in \theta_j^1}\ score_{ij}^1$$
$$childScore_2 = avg_{j \in \omega_2}\ avg_{i \in \theta_j^2}\ score_{ij}^2$$

with $\omega_l = \{j : head\ words\ in\ child\ node\ l\}$,
$\theta_j^l = \{i: tokens\ t_i\ who\ scored\ highest\ for\ a\ head\ word\ p_j^l\}$

Intuitively, each token became associated with the most similar *head word* in each child (*token scores)* and these scores are in a first step averaged for all tokens per *head word* and In a second step averaged per child. The document will then continue its path to the highest *childScore*.

3. <u>*Headword improvement*</u>: Finally, a token t_i replaces its most similar *head word* p_j (*token score)*, in the child node the document will pursue its path, if following condition is fulfilled: t_i represents the tokens, which have passed this node, better than p_j in the sense that the cosine similarity of the center of all tokens, having passed this node, is closer to t_i than p_j.

In short, each document walks through the tree and its path is determined by comparisons with *head words* of each node until a leaf node is reached. Once arrived at a leaf node (*clustering node*), two new *clustering children* are generated and the document will be assigned to one of them through *head word* comparison. Note here, that the new *clustering node children* are only spawned if a minimal number of documents have reached the *clustering* node and if the maximum number of nodes has not been reached (default: 2048). This minimal number of documents is a parameter with default 50. Ultimately, when all documents have been sent to construct the tree, the entire operation is repeated and all documents are again processed one-by-one, so that the *head words* keep improving as long as the sum of all distances of *head words* to their closest tokens have reached a local maximum. A real clustering example is shown in Figure 6.4.

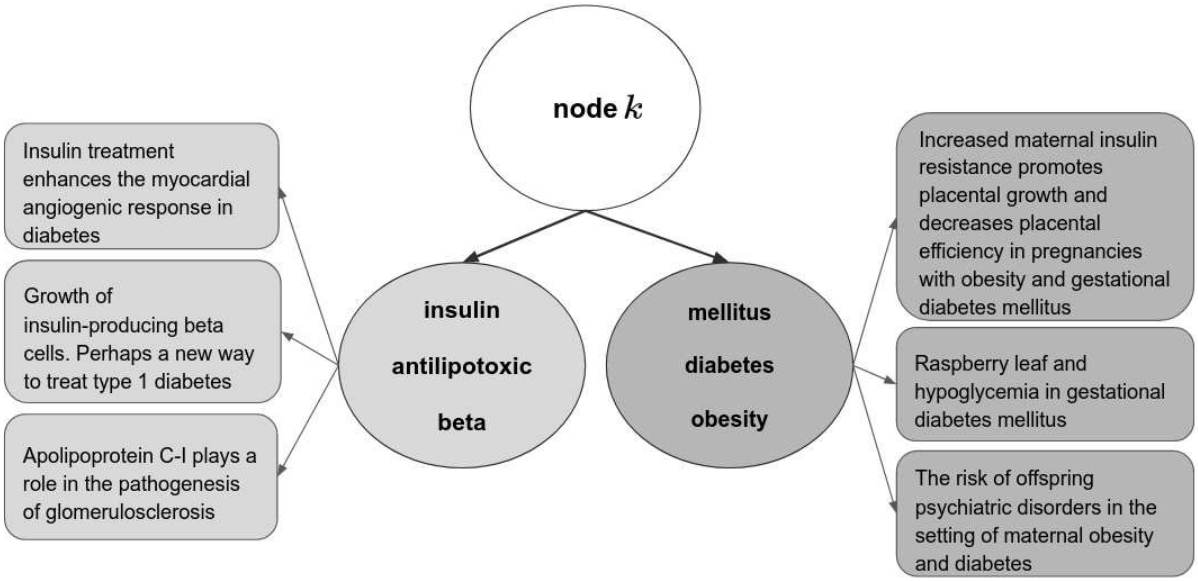


Figure 6.4: Real clustering example of diabetes abstract titles illustrating the *head words* generated on two sub-nodes and three sample abstract titles per node.[294]

A detailed description of the formulas for the hierarchical clustering algorithm, the convergence of the algorithm, and examples are provided in ANNEX 8.

6.2.2.2 Classification

Generally speaking, a *classifier node* symbolises a user-defined topic defined by positive annotations (examples of the user-defined topic) and negative annotations (other documents). An underlying machine learning classifier is trained with those annotations to predict whether a document can be associated with the user-defined topic. In this way, the *classifier node* acts as a barrier and lets documents pass to the underlying nodes, where clustering goes on, only if they correspond to the user-defined topic. By default, Support Vector Machines have been chosen as classifiers, but this configuration is modifiable.

Once the initial tree with root node "*In Scope"*, and underlying *clustering nodes* have been created, a user starts exploring the tree, *via* the interactive interface, and tries to identify a *clustering node* that potentially represents a topic of interest by reading the *head words* and most important documents for each node, compare Figure 6.5. Focusing on such a node provokes the creation of a *classifier node* at the top of the tree; a *clustering child node* under the created *classifier node* and if on this tree level does not already exist a *clustering node* brother, one is created. The *head words* of

the selected node are used as initial positive annotation and the *head words* of the brother node of the selected node serve as initial negative annotations. Still *via* the interface, a user selects proper documents which serve as positive and negative training annotations for the *classifier* (e.g. “Type 2”-Diabetes in Figure 6.5). In the next iteration, at tree rebuilding, each document that enters the tree will first be fed to the *classifier nodes* (e.g. “Type 2”) and if the document relates to “Type 2”, it passes the *classifier node* to the underlying *clustering child node* where clustering continues. Otherwise, the document is redirected to the *clustering brother node*. Compare Figure 6.5 for an overview over the *classifier* node creation.

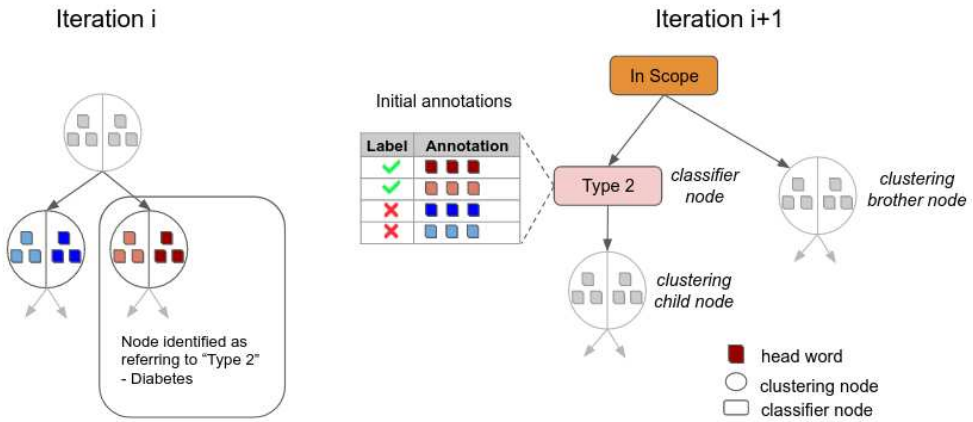


Figure 6.5: *Classifier* node creation with *clustering child node* and *clustering brother node*.

A document will always be compared first to *classifier nodes* to see if it corresponds to a user-defined topic and otherwise goes to the *clustering brother node* to continue being clustered.

Furthermore, in an optimal clustering the *clustering nodes* under a *classifier node* only regroup documents pertinent with the user-defined topic. However, in practice this is not the case, especially in the first iterations, due to the insufficient annotations to train the *classifier nodes* which impacts model performance. See Figure 6.3 iteration 1 for an example where the purity (light red) of some nodes, with respect to the proportion of documents referring to “Type 2”, is shown. For this reason, the user has two possibilities to improve classification performance via the interface:

- correcting misclassification in the lower levels under a *classifier node* by selecting those documents as negative annotations for classifier training
- concentrating on different tree branches which might contain documents related to the user-defined topic, for instance “Type 2”, that were not recognised by the *classifier*. Adding

those documents as positive annotations might help the classifier to recognise those documents in the following iteration.

Throughout this iterative process a user also continues to identify further sub-topics for already created *classifiers*, leading to an own hierarchy of *classifiers* at the top of the tree, compare Figure 6.2.

Iteratively a user continues to create classifiers, choose appropriate documents as annotations for the classifiers, and fix potential misclassifications until the system eventually converges towards a user-desired clustering solution of topics of interest.

6.2.2.3 Interactive interface

The interactive interface visualises the tree structure via nested circles and has been developed in D3, jQuery and JavaScript. For a global view on the interface please see Figure 6.6. The colored nodes represent *classifier nodes.*

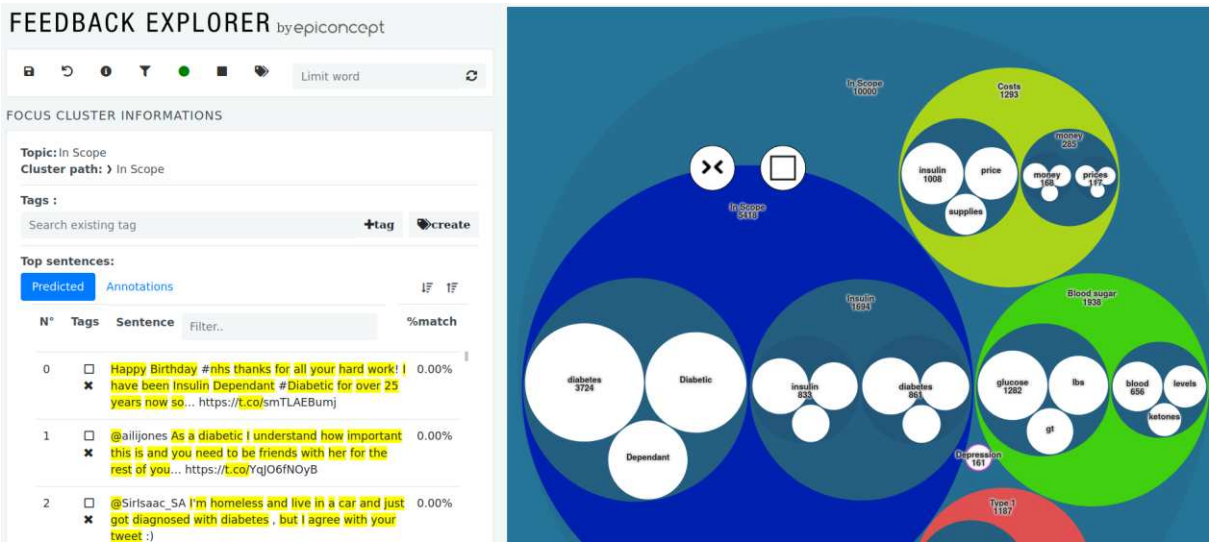


Figure 6.6: Visual user interface overview. Colored circles represent user-defined topics (*classifier nodes)*. *Head words* are shown in white circles in each node.[294]

Furthermore, clicking on a node provokes an automatic zoom on the node and the *head words* of its children as seen in the top-left image of Figure 6.7. For classifier nodes, the interface also provides

information about the positive and negative annotations, as illustrated in the top-right picture. The bottom-left image shows the closest documents to the node ordered by the *childScore* for this node. It is also possible to visualise the least representative documents as shown in the bottom-right image.

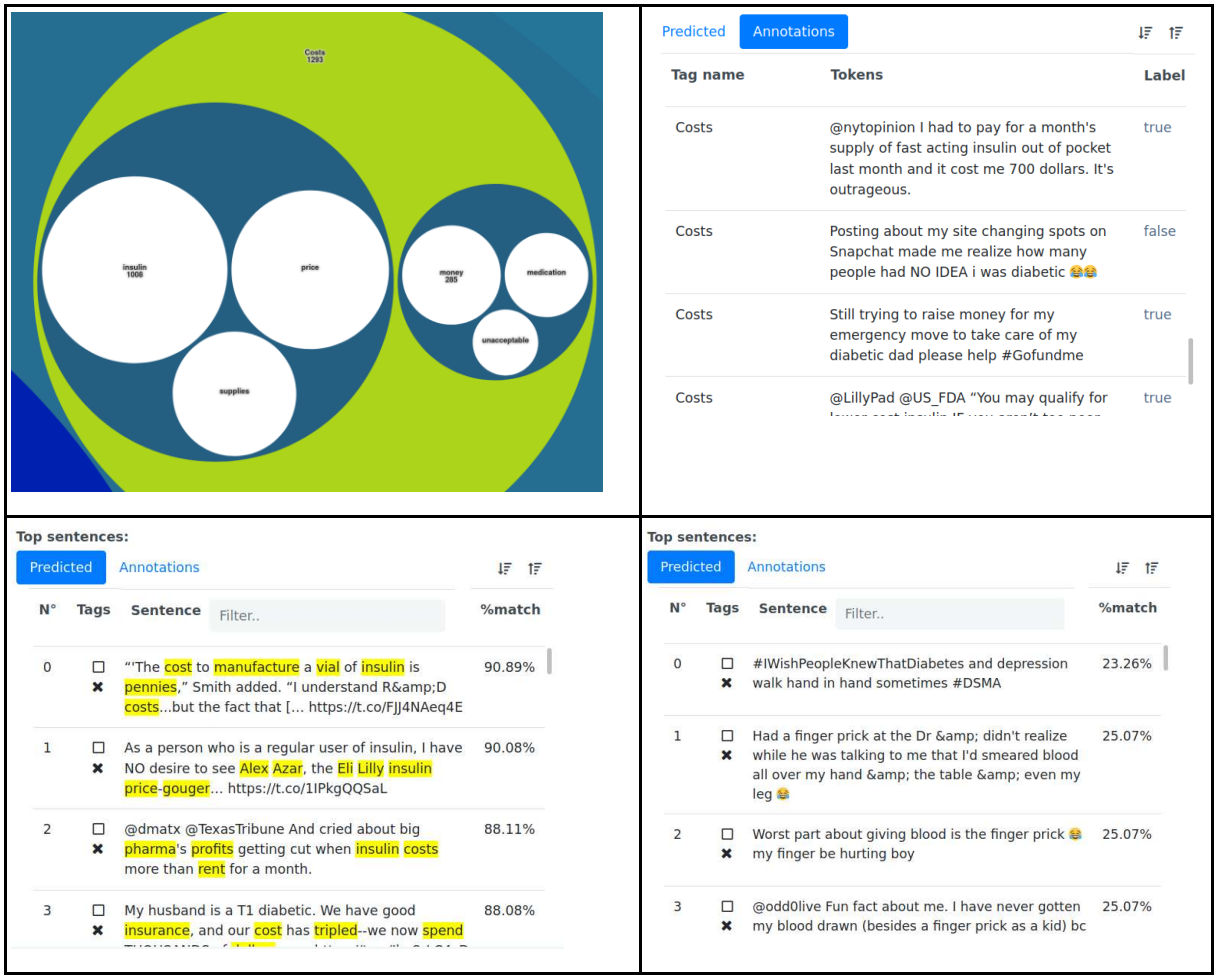


Figure 6.7: Zoom into the *classifier node* "Costs" with the *head words* of its children in the top-left. The top-right shows positive and negative annotations for this *classifier*. The bottom-left image shows the closest sentences and the bottom-right shows the least closest sentences having passed this node.

The user acts on this tree while the system guides him to find a user-desired solution.

6.2.2.4 Active learning

In this work a novel active learning strategy was developed which combines *uncertainty sampling* with the hierarchical tree architecture in order to minimize the user annotation effort and to accelerate convergence towards a user-guided solution.

The idea is to choose the best training annotations, for a given class (MeSH code), by selecting documents from deeper levels of the tree. At a given *classifier node* (ex.: Type 1), a simple *uncertainty strategy* would select the documents the underlying model is most uncertain about (prediction near to 0.5). However, this choice might lead to picking documents from only one specific sub-branch below the *classifier node* and not from the whole variety of sub-branches. For this reason, we hypothesized that picking the documents the model is most uncertain about from the tree level containing most nodes under the *classifier node*, ensures a better distribution in the vector space and in consequence provide more diverse annotations.

A visualisation of this strategy and how we evaluated it is shown in Figure 6.8 for the sample MeSH code *type 1 diabetes*. In the first iteration the tree is built with only two *classifier nodes*: *In Scope* and one *classifier node* for the class to be tested, in this example “Type 1”, which has no training annotations yet. In the next step, the depth level D_{max} in the tree containing most nodes is determined. From each of the nodes on the level D_{max} randomly *batchSize*-many documents (default: 50) are chosen consecutively which constitute the initial annotations to train the “Type 1” classifier. Then, the tree is rebuilt, including the trained *classifier node*, which filters “Type 1” related documents in his sub-tree (*posTree)* and non “Type 1” related documents continue being clustered in the sub-tree under its *clustering brother node* (*negTree)*, see Figure 6.8. In scenarios with imbalanced classes, the *negTree* is usually larger as it concentrates more documents. This is the case here, where the target class is “Type 1” but the majority of documents have a MeSH code not related to “Type 1”. In each subtree the level D_{max} containing most nodes is identified and *batchSize / 2* many documents are selected consecutively, with respect to *uncertainty sampling*. Contrary to the first iteration, for each node on level D_{max} the documents the model is most uncertain about are related to the “Type 1” classifier (prediction ~ 0.5). Those documents are added as new training annotations for the *classifier* and the iterations continue to add more training data. An example of how documents are selected consecutively on a given D_{max} level is included in Figure 6.8.

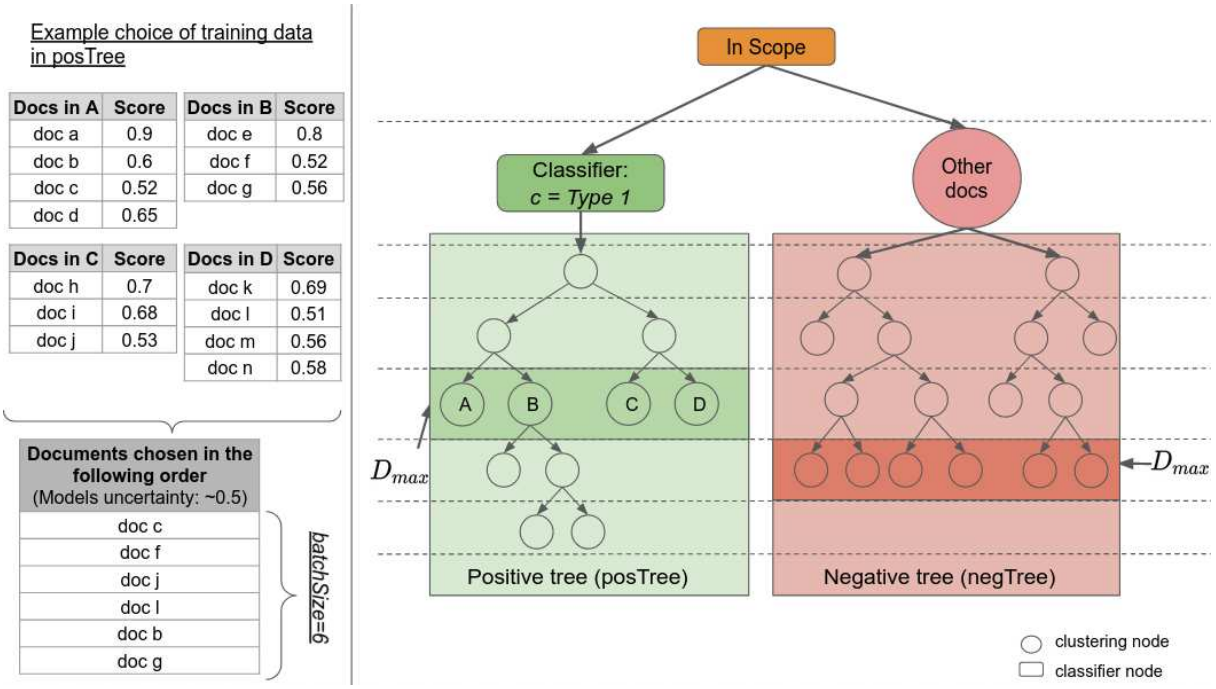


Figure 6.8: Active learning tree. The positive tree is the sub-tree under the *classifier node* "Type 1" and the negative tree is the sub-tree under his clustering brother.[294] On the left side a sample is provided of the document selection process.

Through the interface a user has then the choice to either apply the automatic active learning strategy or to apply the simpler *uncertainty sampling* strategy. The interface contains a functionality allowing users to change the order of the sentence by their scores and thus offering the opportunity to see the documents a node is least confident about.

This methodology was integrated in an open-source tool called *FeedbackExplorer* (FBE). For this reason, in the following we may refer to this methodology as *FeedbackExplorer* or FBE.

6.3 Results

The evaluation of this methodology comprised the comparison of our hierarchical clustering and the active learning strategy with popular, open-source algorithms. Memory consumption is also examined. As a reminder, the goal of this work was not the establishment of a new state-of-the-art performance but rather to demonstrate that near state-of-the-art performance is possible while

addressing limitations of current systems such as usability for non-experts, memory consumption or the lack of interpretability.

6.3.1 Hierarchical clustering

The hierarchical clustering of our methodology is compared with the Hierarchical agglomerative clustering (HAC) algorithm implemented in the popular open-source library *scikit-learn.*[83] We used this algorithm with *complete* linkage criterion. For an equal comparison we ran both algorithms with two configurations, one with 32 leaf nodes and one with 64 leaf nodes. Both algorithms were executed on 50.911 diabetes abstracts as described in the methods section.

We ran FeedbackExplorer's clustering 10 times with random document order due to its streaming character which leads to different clustering solutions for a different order of documents. The HAC algorithm is stable in this regard and one execution was sufficient. Table 6.2 shows the *F1-Score,* specifically designed for hierarchical clustering and introduced in chapter 3.9.2.2. The confidence interval for FeedbackExplorer is given in brackets due to several executions. For both configurations HAC achieved the best performance. Nonetheless the clustering performance of FBE reached a *F1-Score* of 0.73 and 0.74, close to the HAC results.

Tree configuration	**HAC**	**FeedbackExplorer**
32 leaf nodes	0.76	0.73 [0.712, 0.757]
64 leaf nodes	0.77	0.74 [0.717, 0.760]

Table 6.2: *F1-Score* for Scikit-learn's HAC and FeedbackExplorers hierarchical clustering.[294] In parentheses the confidence intervals

6.3.2 Active learning

Four strategies were compared to assess active learning performance:

- *Random strategy*: The algorithm chooses documents randomly to train the classifier.
- *Uncertainty strategy*: The model chooses the documents it is most uncertain about, in this case when the model prediction for a given class is close to 0.5.

- *FeedbackExplorer* strategy: The proposed strategy which combines uncertainty sampling with exploiting the tree structure.
- *CNN-Zhang*: This strategy was introduced by Zhang et al. and combines convolutional neural networks (CNN) with the active learning strategy *expected gradient length* to classify documents.[84,318] This strategy picks documents if they contain words that are likely to most affect the word embeddings by calculating the *expected gradient length* with respect to the embeddings for each word.[319] The authors provided the code of this method on GitHub.

Performance was assessed for all diabetes-related MeSH codes listed in Table 6.1 individually. For a given MeSH code, all associated documents were considered the positive class while all other documents were considered the negative class. This led to highly imbalanced datasets for most MeSH codes. For this reason, it was also interesting to consider the number of positive annotations which have been identified by the different active learning strategies. Typically, active learning is evaluated on how the performance evolves with more and more documents on the first iterations. A random subset of 2,000 documents was chosen, containing at least, if possible, 50 documents per MeSH code to ensure the presence of each code and thus the ability to train the classifiers. The 2,000 documents were randomly split into 1,000 training and 1,000 test documents. Performance is evaluated for 50, 100, 150 and 200 annotations to train the classifiers per strategy.

Table 6.3 indicates the weighted average performance over all MeSH codes.

# train. data	Random				Uncertainty Sampling				FeedbackExplorer				CNN-Zhang			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
50	0,87	0,62	0,57	0,51	0,83	0,56	0,60	0,50	0,88	0,63	0,44	0,49	0,81	0,24	0,31	0,20
100	0,86	0,62	0,51	0,49	0,88	0,68	0,64	0,62	0,90	0,71	0,51	0,56	0,86	0,39	0,59	0,42
150	0,88	0,68	0,46	0,47	0,90	0,75	0,62	0,63	0,90	0,75	0,59	0,60	0,88	0,52	0,72	0,55
200	0,89	0,62	0,43	0,45	0,91	0,77	0,53	**0,61**	0,91	0,71	0,58	**0,62**	0,90	0,58	0,79	**0,63**

Table 6.3: Weighted average of Active learning performance over all MeSH codes.[294]

After 200 training instances the performance is similar for the three non-random strategies. However, this might be misleading as these averaged values hide important variations of these models depending on the MeSH they consider. Particularly, MeSH codes with few relevant documents lead to low performances in general.

ANNEX 9 provides a detailed overview over the performance of all MeSH codes. In most cases FeedbackExplorer's and Zhang's CNN's performance was similar after training with 200 instances.

But for some MeSH codes FeedbackExplorer achieved highest performance (ex.: D048909: Diabetes complications, D003928: Nephropathies or D005320: Fetal Macrosomia) and likewise for some MeSH codes Zhang's CNN performed better (ex.: D003925: Angiopathies, D003930: Retinopathy or D016640: Gestational Diabetes).

Pos. class	#	Random						Uncertainty Sampling						FeedbackExplorer						CNN - Zhang					
		pos	neg	Acc	Prec	Rec	F1	pos	neg	Acc	Prec	Rec	F1	pos	neg	Acc	Prec	Rec	F1	pos	neg	Acc	Prec	Rec	F1
Complications D048909	50	31	19	0.79	0.77	0.94	0.84	29	21	0.71	0.70	0.90	0.79	29	21	0.78	0.88	0.72	0.80	26	24	0.62	0.72	0.78	0.59
	100	67	33	0.73	0.69	0.98	0.81	39	61	0.78	0.77	0.91	0.83	61	39	0.83	0.88	0.82	0.85	57	43	0.73	0.81	0.74	0.76
	150	94	56	0.81	0.79	0.93	0.85	57	93	0.83	0.84	0.87	0.86	88	62	0.84	0.84	0.90	0.87	76	74	0.78	0.83	0.83	0.81
	200	125	75	0.81	0.79	0.93	0.85	78	122	0.85	0.91	0.82	0.86	110	90	0.85	0.85	0.91	**0.88**	104	0.96	0.81	0.84	0.85	0.84
Test set: 591 (pos) / 409 (neg) || Train set: 557 (pos) / 443 (neg)																									
Angiopathies D003925	50	14	36	0.86	0.78	0.48	0.60	11	39	0.84	0.86	0.30	0.45	17	33	0.88	0.80	0.53	0.64	10	40	0.79	0.02	0.33	0.03
	100	23	77	0.86	0.93	0.37	0.53	55	45	0.90	0.76	0.73	0.75	45	55	0.90	0.81	0.56	0.72	43	57	0.85	0.35	0.87	0.52
	150	35	115	0.85	0.93	0.33	0.48	76	74	0.88	0.67	0.86	0.75	65	85	0.89	0.75	0.75	0.75	69	71	0.87	0.54	0.85	0.63
	200	44	156	0.86	0.94	0.35	0.51	87	113	0.87	0.84	0.46	0.60	76	124	0.90	0.87	0.61	0.72	95	105	0.90	0.68	0.83	**0.74**
Test set: 209 (pos) / 791 (neg) || Train set: 178 (pos) / 822 (neg)																									
Cardiomyopathies D058065	50	1	49	0.95	0.24	0.27	0.25	3	47	0.97	0.5	0.03	0.06	1	49	0.97	0	0	0	2	48	0.97	0	0	0
	100	3	97	0.97	0	0	0	15	85	0.97	0.56	0.47	0.51	3	97	0.97	0	0	0	3	97	0.97	0	0	0
	150	5	145	0.97	0	0	0	23	127	0.97	0	0	0	5	145	0.97	0	0	0	7	143	0.95	0.02	0.10	0.03
	200	7	193	0.97	0	0	0	26	174	0.97	0	0	0	13	187	0.97	0	0	0	12	188	0.97	0.01	0.20	**0.02**
Test set: 30 (pos) / 970 (neg) || Train set: 28 (pos) / 972 (neg)																									

Table 6.4: Active learning performance for all four strategies.[294]

Specific results are highlighted in Table 6.4 on three MeSH codes. In addition, information about the number of positive and negative training instances are provided. For the MeSH code *diabetes complications (D048909)* FeedbackExplorer reached highest performance after 200 training instances with a F1-Score of 0.88, while Zhang's method achieved best performance for *diabetes angiopathies (D003925)*. The last MeSH code *diabetes cardiomyopathies (D058065)* was an example for a MeSH code with poor results for all strategies due to the few positive training examples. The occurrence of some MeSH codes with low predicted performance, mainly due to too few positive examples in the training data, is one reason why the overall performance measures in Table 6.3 were not higher.

6.3.3 Memory consumption

An overview of memory consumption in MegaBytes (MB) and the execution time in minutes (min) is displayed in Figure 6.9. Scikit-learn's HAC algorithm and FBE were both executed for 10.000, 20.000, 30.000, 40.000 and 50.000 documents. We noticed that the memory consumption for FBE stays almost constant with the growing number of documents while the HAC algorithm grows exponentially. However, the gain in memory consumption equates to a longer running time.

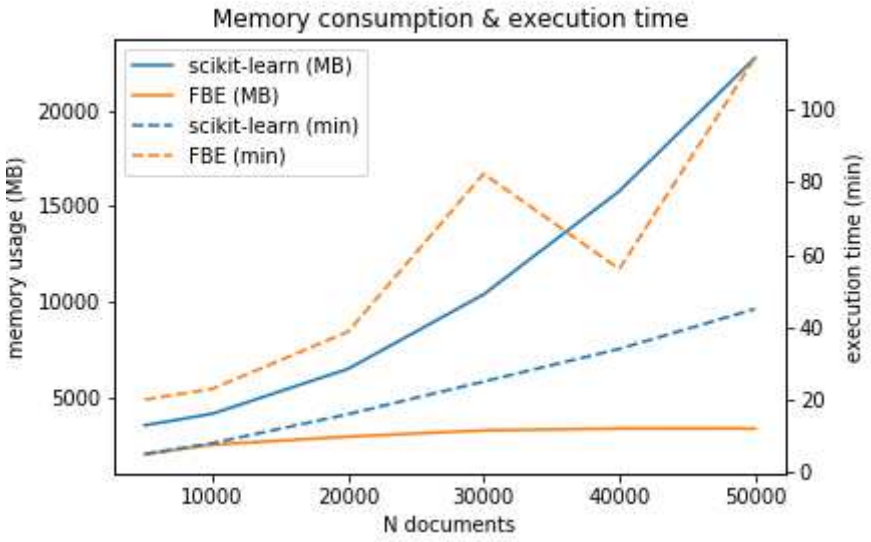


Figure 6.9: Memory consumption in Megabytes (MB) and execution times in minutes (min) .[294]

6.4 Discussion

6.4.1 Principal results

In this work, a visual interactive user interface has been implemented, enabling users without computer science knowledge to discover and target topics of interest in biomedical text data to improve the literature exploration in evidence-based clinical decision making. A novel hierarchical clustering algorithm groups documents in an interpretable fashion using *head words*. There exist two possibilities to minimize the data annotation effort. Either through manual intervention via the

interface by picking pertinent documents the model is most uncertain about; or applying the proposed active learning strategy which combines *uncertainty sampling* with the hierarchical tree form.

Our methodology reached near state-of-the-art performance when compared with efficient hierarchical clustering algorithms and various active learning strategies. Beyond the performance, FeedbackExplorer addresses several existing limitations in common machine learning algorithms to extract information from textual data such as the struggle of injecting domain knowledge to the model; the requirement to define the desired number of clusters *a priori*; the interplay of classification and clustering in one methodology or the challenge of applying advanced artificial intelligence methodologies for users without computer science knowledge. Those functionalities make it an attractive asset for the analysis of laboratory results, electronic health records, publications or social media data by health professionals. Furthermore, we demonstrated that our approach is memory efficient and stays stable with increasing documents making it ideally suited for handling large datasets. Through heavier parallelisation of the underlying *Spark* framework the rising execution time could be minimized.

This methodology will be of particular interest specialists who need a quick overview of the literature concerning a specific theme.

6.4.2 Comparison with prior work

Different general-purpose natural language processing systems have been proposed to extract information from clinical text data. The review of Wang et al.[320] identified the most frequently used systems as: the Clinical Text Analysis and Knowledge Extraction System (cTAKES),[321] MetaMap[322] and Medical Language Extraction and Encoding System (MedLEE).[323] These systems have been deployed on a wide range of information extraction tasks based on electronic health records such as the identification of respiratory findings,[324] the detection of smoking status,[325] or pharmacovigilance.[326] However, these systems have difficulties incorporating machine-learning models.[327] Zheng et al. tested several clinical NLP systems and found that they were challenging to set up and configure which led to a general dissatisfaction and thus hampers wide adoption.[328] Soysal et al. addressed the frequent issue of having to customize systems for existing tasks, which frequently requires substantial NLP knowledge, by proposing the NLP toolkit Clinical Language Annotation, Modeling, and Processing (CLAMP) enabling users to build customized NLP pipelines user a user-interface.[327]

But as CLAMP was trained on specific corpora to extract specific types of information, adopting this system beyond its original purpose might harm performance and generalizability.[329] Montani and Striani emphasized in a recent survey the importance of transparency and explainability, of how conclusions are obtained, in AI powered clinical decision support systems.[330] This is in line with our methodology of directly involving users and creating a solution in synergy with the system. Ozaydin advocates as well that including health professionals more in the model development and evaluation helps shed light on the black box systems.[331] An original approach is the voice-enabled chatbot "Plutchik" which is able to search in medical databases, retrieve and communicate medical information.[332] Yet, more sophisticated machine and deep learning methods are not yet implemented.[332]

This methodology aligns with Shneiderman's visualization mantra "*Overview first, zoom and filter, then details-on-demand"* when designing graphical user interfaces.[333] A first *overview* is provided through the visualisation of the top level nodes; *zooming and filtering* is achieved by clicking on specific nodes which provides information about the *head words* of the node; and *details-on-demand* can be obtained by reading the corresponding documents of the targeted node.

To the best of our knowledge, FeedbackExplorer was the first decision support system combining topic exploration, topic targeting with a user-friendly interface and minimized memory consumption and annotation effort in a single methodology. This work enables health professionals to rapidly gain insights into clinical text data to improve decision making.

6.4.3 Strength & Limitations

A crucial strength of the proposed methodology is that non-experts with no programming knowledge are able to explore and target topics of interest in an unstructured textual corpus *via* an interactive and user-friendly interface. Transparency is profoundly improved through *head word* and tree structure visualisation. This is in line with Alcacena et al. who suggested that proper visualization can increase the transparency of machine learning.[85] Transparency in clinical decision support systems is key to ensure adoption by clinicians.[86] Through the interaction between human and system interpretability is increased as has been shown by Lage et al.[87] Our approach is memory efficient due to the streaming character and thus able to be used on computers with limited memory. In addition, the system is built upon the large-scale data processing framework Apache

Spark, allowing fast execution time and the capability of tackling large datasets. A useful feature, as the analysis of large text corpora usually is quite computation intensive.[334] The possibility to dig into topics, when an interesting cluster is found combined with an interpretable result in the form of *head words*, allows to identify sub-themes and to visualize knowledge in such a structured hierarchical format, especially for rare events or uncommon events in clinical practice. Moreover, the data annotation effort is minimized through the proposed active learning strategy by picking the most impactful data instances. The parameter defining the desired number of clusters to be obtained is not required with our methodology contrary to many classical clustering algorithms. Injecting potential domain knowledge to drive topic extraction is still a challenging task. We proposed a way to use domain knowledge to specifically search for topics of interest and test hypotheses to improve the clinical decision making. This may be particularly interesting in the field of rare diseases, where clinical practice based on valid evidence is challenging.[335] Besides, the dynamic nature of our approach allows us to add more documents with time without a complete retraining, allowing the evolution of topics of interest to be studied. A last advantage is that our methodology can be adapted to different languages by providing the respective word embeddings that are easily found on the web.

Concerning the limitations, a constraint of our methodology is the manual interaction which prevents the creation of a huge number of classifiers. However, usually a user has a priori knowledge of the domain and is able to broadly target the desired clusters which avoids having to define a large number of classifiers. Another limitation is that our methodology was only evaluated on a single dataset. A future direction of investigation is the validation of the methodology on other datasets to ensure generalization and portability in other contexts. Due to the streaming nature, the constructed tree depends highly on the order of the documents, which might affect data interpretation. Therefore, in a future work the stability of the hierarchical clustering regarding the document order should be further improved by for instance aggregation techniques. In some cases we observed oscillations in the active learning performance in the sense that performance is not always improving with more training data, a phenomenon frequently observed in the active learning field.[336] A future direction to explore should be the evaluation of the approach on a sample of end-users of different profiles and level of expertise in clustering techniques. Another improvement on the tree creation could be the implementation of further features such as merging different tree branches if they correspond to similar concepts.

6.4.4 Conclusion

An interactive user-interface for users without computer or data science knowledge for the exploration of unstructured biomedical text data as clinical decision support has been proposed. Through a semi-guided, active participation of the user and the visualisation of *head words*, the algorithm converges towards a user-defined solution while improving transparency. Various advantages are combined in a single methodology from leveraging domain knowledge to target topics of interest; over minimizing the manual annotation effort exploiting active learning resulting in a quicker convergence; to reducing memory consumption through data streaming and thus enabling the handling of large textual corpora thanks to Spark's parallelism abilities. We demonstrated that the combination of all those beneficial features led to near state-of-the-art performance. The developed system might be beneficial for healthcare professionals with limited computer science knowledge to quickly gain an overview of specific topics to ultimately improve the literature exploration in the clinical decision making process.

CHAPTER VII: Discussion and Perspectives

This thesis explored how artificial intelligence methods benefit diabetes research. The three principal sections in this thesis constituted: 1) the identification of diabetes concerns and diabetes-related topics of interest in social media data (chapter IV); 2) the detection of cause-effect relationships in diabetes-related tweets (chapter V); 3) and the development of a methodology to explore textual data and target user-desired topics, for healthcare professionals with limited computer science knowledge, via an interactive user interface (chapter VI). Hereafter, the principal results of these three studies are discussed, followed by an overall synthesis and conclusion of this thesis. The final section elaborates on potential future investigations and open challenges.

7.1 Principal findings

In the first objective, an innovative preprocessing pipeline was developed to focus on tweets containing personal and emotional language, geolocalized from the US, allowing us to investigate diabetes-related topics and associated emotions. A main concern that emerged was the theme of insulin pricing, including access to insulin and the identification of financial sources for insulin such as 'glucose guardians' or donations. This theme was frequently accompanied with negative emotions, which confirms the burdensome situation of high insulin prices in the US.[262]

However, positive emotions related to the theme of insulin pricing were observed concerning the fight for affordable insulin in the community. Regarding insulin pricing, this was the first work to demonstrate and quantify the crisis in the USA of people with or talking about diabetes based on social media data.

Moreover, fear, anger and sadness were expressed when it comes to diabetes-related complications and comorbidities as well as substantial frustration about people's inability to differentiate type 1 and type 2 diabetes. Positive emotions were observed in tweets referring to support and solidarity among the diabetes online community validating the positive character of the diabetes online community.[146]

Furthermore, we noticed associations of insulin pricing topics to be more frequent in cities with high mean incomes.

From a methodological point of view, thanks to a machine learning-based preprocessing pipeline, we were able to exclude most noisy tweets and focus only on relevant tweets, namely personal, non-joke and emotional. Many other studies do not make the effort of such a strict preprocessing pipeline.[219,337,338] Another advantage of leveraging machine learning algorithms was that we were able to analyze a large number of tweets with a great variability. This question of representativeness has to be considered with caution, as the variability in people with diabetes on Twitter does not inevitably mirror real-world diabetes patients. But this lack of recall is counterbalanced by the vast volume of data analyzed, thanks to ML algorithm.[81] However, relying on machine learning algorithms also introduced confirmation bias due to manual labeling. As the performance of the algorithm is not flawless, there were still irrelevant tweets in our analyses. With the complexity in human language and text data removing noise certainly is an endless challenge. Moreover, we developed an innovative geolocation engine which exploited several meta-data fields (*place full name, user location, user description*) and thus were able to infer the city-level geolocation for 63% of our preprocessed tweets. Contrary to other studies which exploited only the *user location* field[229] or geolocated only on a state-level.[149]

Throughout the second objective, we demonstrated the feasibility of inspecting both explicit and implicit multi-word causal information and cause-effect relationships, whereas most former work studied the simplified task of explicit causality.[80,269,290] The preprocessing pipeline was aligned with the first objective. The transfer learning paradigm was adopted to fine-tune the pre-trained language model *BERTweet* to detect sentences containing causal information in a first step. In a second step a single conditional random field layer with *BERTweet*-based and discrete features outperformed fine-tuning *BERTweet* architectures in the *cause-effect* relationship detection task. We were able to extract cause-effect patterns in 20% of the preprocessed sentence, contrary to other works which had focused only on explicit causality and extracted less than 2% causal patterns.[80] The obtained performance measurements were satisfying given the challenging task and the imbalanced dataset, but could certainly be improved with more data and better data quality. The lack of recall is again compensated by the massive amount of data, whereas the lack of precision was compensated through the semi-clustering approach. An interactive visualisation of the *cause-effect* network has been provided for interested readers to explore these relationships. The largest cluster, playing mainly the role of a cause, was “Diabetes” which most frequently led to the effects “death” and “fear”. Another large cluster was “Death” being frequently caused by various factors referring to insulin pricing. This confirms the strong presence and importance of the topic of insulin

pricing in the US diabetes community, which was already detected during the first objective. Besides, regular clusters were “Type 1 diabetes” causing regularly “insulin pump”, “hypoglycemia”, “sickness”, “finance” and emotion-related effects; the cluster “Insulin” was strongly related to “sickness”, “medication”, “finance”, “death” and emotions.

The third objective addressed existing limitations in literature exploration to improve the evidence-based clinical decision making process. For this purpose, a visual user interface has been developed allowing users without a computer science background to iteratively explore, discover and target topics of interest within a given corpora. The main structure of the algorithm is shaped by a hierarchical clustering which structures documents in a streaming fashion. A key design choice was to define the clustering tree nodes via *head words* which should improve interpretability. By putting a human-in-the-loop, a solution is found hand-in-hand with the machine rather than relying on a stand-alone blackbox model. As a consequence, not only domain knowledge is incorporated into the system, but also transparency is enhanced. With these design choices, adoption by health professionals should be facilitated. Human interaction is minimized through an active learning strategy, leveraging the hierarchical skeleton, by proposing relevant documents that the model is uncertain about to be annotated. As a consequence of human interaction, there is a limit to the number of classifiers that can be created.

The aim of this methodology was not to establish a new state-of-the-art performance, nevertheless we have shown to reach near state-of-the-art performance. We demonstrated how frequent limitations in machine learning algorithms to extract information from text data can be tackled. Contrary to many clustering or topic models our approach is independent from defining a fixed number of clusters prior to the training. Advanced machine learning algorithms are made accessible to users without programming skills. Moreover, we made a contribution towards memory efficient modeling with a stable memory consumption with an increasing number of documents. The resulting longer execution time for more documents can be balanced through heavier parallelisation of the underlying Spark Framework.

In addition, the ability to zoom into identified topics, thanks to the hierarchical structure, enabling the exploring of sub-topics, makes the methodology even more appealing for health professionals.

7.2. Synthesis and conclusion

7.2.1 Epidemiological contributions

The feasibility of exploiting social media data to capture real-life, patient reported emotions and concerns has been demonstrated which represents a cost- and time-effective way of augmenting psychosocial and epidemiological research. Using social media posts we were able to capture emotions, and concerns of people with or talking about diabetes, including diabetes distress related topics that were not included in current scales to evaluate diabetes distress. These data reflect the online user behavior as well as the information they are exposed to. Besides, we demonstrated the practicability of extracting both explicit and implicit cause-effect relationships exploiting modern machine learning architectures and natural language processing. This offers the possibility to enhance our knowledge of diabetes in a real-life setting exploiting patient-reported outcomes. Future studies should be encouraged in considering social media data as complementary data source. This is in line with Gabarron et al. who showed that intervention studies using social media appear to have positive impacts on health outcomes in T1D patients and could be beneficial for people with type 2 diabetes as well.[339]

Public debate about diabetes issues could be promoted using Twitter data and thus have implications for health policy makers, health promotion practitioners and clinical decision making. Social media data will serve in developing policies and interventions that include key concerns in people with diabetes to ultimately improve health outcomes. For this to become reality, our findings require validation in epidemiological studies.

In addition, currently there is a lack of adoption of efficient tools to analyze and extract appropriate information out of the exponentially growing textual data due to: poor user-interfaces; the failure to integrate those tools naturally in the health care professional's routine; the design of those tools from a too strong technical perspective or a lack of physician acceptance or black-box models.[49–51]

To tackle these issues, we have proposed an interactive user-interface for users without computer science and programming knowledge to explore unstructured clinical text information to ultimately improve the literature summarization part in the clinical decision making process. Transparency and interpretability is ensured through actively involving the user in the solution finding process and *head word* visualisations. The ability to actively interact with the model cultivates trust within health

practitioners.[310] Nevertheless, more time needs to be invested to further improve the tool; to test it on a sample of end-users to include their feedback and confirm portability on other datasets.

7.2.2 Technical contributions

This thesis mirrors well the evolutions in natural language processing. The first objective addressed diabetes concerns on Twitter based on *FastText* embeddings. With the advent of context-aware word embeddings, namely *BERT,* during this thesis, we also employed these more powerful embeddings in the second objective to tackle causality.

Besides, this work was also a playground for testing different machine learning and natural language processing methods and to validate them for the analysis of social media data. Generally, the foundation in any machine learning classifier is its training data, which defines the ground truth of any signal which is inferred by the classifier on unseen data. In consequence, the data quality largely determines the classifier outcome. We have shown that active learning is an efficient way to increase the training data with discriminative samples to minimize the annotation effort and ideally increase performance.

Furthermore, with this thesis we have highlighted that, indeed, machine learning methods can be exploited for diabetes epidemiology at all stages of the data pipeline: preprocessing data, geolocating data, filtering data, predicting and clustering of information. Globally, it enabled us to target larger datasets and thus greater variability in the data and stronger representativeness.

With the development of our clinical decision support tool for literature and text exploration we tackled various technical bottlenecks ranging from limitations of current NLP methods, such as specifying the desired number of topics beforehand or weak scalability of topic models, over improving the annotation effort through active learning to minimize memory consumption thanks to the streaming nature and underlying Spark framework. Automatically finding a clustering solution without human intervention remains a challenging problem. Human involvement in model training represents an alternative until more powerful machine learning algorithms, which ensure transparency and interpretability, will be developed. Within AI applications there is always a trade-off between model performance and transparency.[88] Simple models such as decision trees are generally interpretable but tend to sub-optimal performance, whereas complex models such as deep neural networks deliver oftentimes superior performance, but interpreting the main factors that determine performance can be burdensome.[88] With this methodology we contributed towards

the goal of more explainability in healthcare, by consciously deciding against neural network based architectures and proposing a modular system in which simpler ML models are interchangeable.

While an advantage of using social media is that it does not suffer from the same biases like traditional epidemiology, it is fair to acknowledge that social media undergoes different biases.
A selection bias is introduced as Twitter is used by only a part of the population, in the case of the USA 23% of all adults use Twitter, raising awareness for the representativity.[19] Our analyses were restricted to a single social media platform, leaving out relevant information and demographics present on other platforms.[23] This platform bias could be addressed in a future work. Besides, we are subject to reporting bias, as we can only analyze what users report and due to social stigma that certain topics might not be heard in fear of public backlash.[96] Another important bias is the algorithmic bias, which represents systematic prejudice due to erroneous assumptions incorporated into the machine learning algorithms.[340] In particular, algorithmic bias has been observed in word embeddings and language models, in form of gender bias and social discrimination.[341,342]
Confirmation bias or human bias occurred during data labeling as we unconsciously treat data in ways that affirm preexisting opinion, beliefs and hypotheses.
Activity bias comes from the fact that most users are silent spectators and mostly watch the content of others, while a minority generates most content and thus dominates with their opinions.[343]

Moreover, this thesis was realized entirely based on the open science spirit. On the one hand we used existing frameworks (PyTorch, Apache Spark) and programming languages (Python, Scala, JavaScript) with associated packages (scikit-learn, gensim, D3, etc.). On the other hand, we open-sourced our algorithms under https://github.com/WDDS/. By openly sharing AI algorithms we contributed to computational transparency and interoperability.[88] We strongly encourage interested researchers to validate them and further extend them for their purposes.

7.3 Research perspectives

This thesis explored new data sources, developed and improved machine learning algorithms and investigated existing limitations at the intersection of AI and healthcare. Even if this work sheds light on the question how social media can be leveraged for diabetes epidemiology, various new doors have opened with remaining research perspectives to be addressed. Some of these potential directions to explore are outlined in the following.

7.3.1 Extension to further countries

This thesis highlighted the importance of social media data as complementary data sources for diabetes research. To establish the preprocessing and analysis algorithms we centered our work on tweets from the USA. A natural extension is to translate these algorithms to other countries and regions. This would enable comparisons between countries and the identification of specific topics of interests and concern.

It is a challenging task to represent as many countries as possible, given that in some African and Southern Asian countries, an important part of the population is still not connected with the internet.[17]

Word embeddings exist in various languages and can easily be downloaded. Currently the training data for the classifiers is only available in English, requiring an adaptation to other languages. One would need to relabel new data or experiment with sequence-to-sequence models or *BERT* models to translate the existing labeled data to the target domains, which could induce another bias.

7.3.2 Crossing socio-economic factors with social media data

We have already started to cross socio-economic factors, such as the household mean income, with social media data, in a small extension of the topic extraction work. However, this subject merits more attention. More socio-economic factors, such as the gini-coefficient, education levels or age and environmental factors such as pollution should be included. An interesting approach could be the study of nested models given the nested structure: a geolocalized tweet is included in a geolocalized city of the tweet, which itself is included in a region, and the region is included in the

country. However, in this context it is important not to fall in the ecological fallacy trap and to attribute measured characteristics on aggregated level (e.g. city mean household income) to an inferior unit (e.g. individuals who tweet and live in the city).

7.3.3 Studying dynamics in social media

In this work static clustering in social media was performed. Few works addressed dynamic clustering or dynamic topic modeling in which the model is able to eject vanishing topics but also include new emerging topics. This could be particularly interesting to capture sudden trends or new disease appearances. A long-term goal could be the creation of a real-time monitoring system tracking tweets in - almost - real-time. This is the concept of the World Online Health Observatory, currently under development at the Luxembourg Institute of Health.

7.3.4 Validation in a traditional setting

Following a data-driven approach in this thesis, we extracted topics and information from a vast amount of health data in a first step. This information can be used to generate new hypotheses which can then be tested in a more traditional setting (e.g. cohort study) in a second step. This form of validation is important to be actionable for public health as we have shown that diabetes-related Twitter data is not representative for all people with diabetes.

7.3.5 Exploring data efficient methods to extract causal patterns

A new approach was presented to identify explicit and implicit causal information expressed in diabetes-related tweets. Due to the fact that causes and associated effects could cover a broad range of entities regarding diabetes, it was challenging to create an appropriate dataset of high quality. Future work could pursue the creation of a larger dataset with superior data quality. Ideally several professionals should participate in the labeling task due to the ambitious and sometimes ambiguous and subjective task of labeling cause-effect associations. Furthermore, more sophisticated model architectures could be investigated with the aim of increasing the generalization capability from a relatively small training set to a large number of documents in the target domain.

7.3.6 Improving the clinical decision support tool

The development of a sophisticated clinical decision support tool is time-consuming and requires an iterative development cycle including the feedback of health practitioners. A next step should be the evaluation by a sample of end-users to include specific needs, correct malfunctions and improve the tool in collaboration with health professionals. Moreover, the algorithmic part of the methodology could be improved including an investigation of the process of convergence and the stability of the tool.

REFERENCES

1. International Diabetes Federation. IDF Diabetes Atlas, 9th edn. Published online Brussels 2019. https://www.diabetesatlas.org
2. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. *The Lancet*. 2018;391(10138):2449-2462. doi:10.1016/S0140-6736(18)31320-5
3. Katsarou A, Gudbjörnsdottir S, Rawshani A, et al. Type 1 diabetes mellitus. *Nat Rev Dis Primer*. 2017;3(1):1-17. doi:10.1038/nrdp.2017.16
4. DeFronzo RA, Ferrannini E, Groop L, et al. Type 2 diabetes mellitus. *Nat Rev Dis Primer*. 2015;1(1):1-22. doi:10.1038/nrdp.2015.19
5. Fletcher B, Gulanick M, Lamendola C. Risk Factors for Type 2 Diabetes Mellitus. *J Cardiovasc Nurs*. 2002;16(2):17-23.
6. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. *The Lancet*. 2017;389(10085):2239-2251. doi:10.1016/S0140-6736(17)30058-2
7. McIntyre HD, Catalano P, Zhang C, Desoye G, Mathiesen ER, Damm P. Gestational diabetes mellitus. *Nat Rev Dis Primer*. 2019;5(1):1-19. doi:10.1038/s41572-019-0098-8
8. Skinner TC, Joensen L, Parkin T. Twenty-five years of diabetes distress research. *Diabet Med*. 2020;37(3):393-400. doi:10.1111/dme.14157
9. Fisher L, Polonsky WH, Hessler DM, et al. Understanding the sources of diabetes distress in adults with type 1 diabetes. *J Diabetes Complications*. 2015;29(4):572-577. doi:10.1016/j.jdiacomp.2015.01.012
10. Polonsky WH, Anderson BJ, Lohrer PA, et al. Assessment of diabetes-related distress. *Diabetes Care*. 1995;18(6):754-760. doi:10.2337/diacare.18.6.754
11. Polonsky WH, Fisher L, Earles J, et al. Assessing psychosocial distress in diabetes: development of the diabetes distress scale. *Diabetes Care*. 2005;28(3):626-631. doi:10.2337/diacare.28.3.626
12. Fenwick EK, Rees G, Holmes-Truscott E, Browne JL, Pouwer F, Speight J. What is the best measure for assessing diabetes distress? A comparison of the Problem Areas in Diabetes and Diabetes Distress Scale: results from Diabetes MILES-Australia. *J Health Psychol*. 2018;23(5):667-680. doi:10.1177/1359105316642006
13. Perrin NE, Davies MJ, Robertson N, Snoek FJ, Khunti K. The prevalence of diabetes-specific emotional distress in people with Type 2 diabetes: a systematic review and meta-analysis. *Diabet Med J Br Diabet Assoc*. 2017;34(11):1508-1520. doi:10.1111/dme.13448
14. Sturt J, Dennick K, Due-Christensen M, McCarthy K. The detection and management of diabetes distress in people with type 1 diabetes. *Curr Diab Rep*. 2015;15(11):101. doi:10.1007/s11892-015-0660-z
15. Fotis J. The use of social media and its impacts on consumer behaviour: The context of holiday travel. Published online 2015.
16. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media. *Bus Horiz*. 2010;53(1):59-68. doi:10.1016/j.bushor.2009.09.003
17. Kemp S. *Digital 2021 April Global Statshot Report*. DataReportal; 2021. Accessed July 15, 2021. https://datareportal.com/reports/digital-2021-april-global-statshot
18. Pew Research Center. Social Media Fact Sheet. Published 2021. Accessed June 28, 2021. https://www.pewresearch.org/internet/fact-sheet/social-media/
19. Pew Research Center. Social Media Use in 2021. Published 2021. Accessed July 15, 2021.

https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/
20. Sayce D. The Number of tweets per day in 2020. Published 2020. https://www.dsayce.com/social-media/tweets-day/
21. Twitter. Twitter Privacy Policy. Published July 16, 2021. Accessed July 16, 2021. https://twitter.com/en/privacy
22. Pew Research Center. Twitter users are younger, more highly educated and wealthier than general public. Published 2019. Accessed October 26, 2021. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/pdl_04-24-19_twitter_users-00-04/
23. Tufekci Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM 14 Proc 8th Int AAAI Conf Weblogs Soc Media*. Published online 2014:10.
24. Fox S. The Social Life of Health Information, 2011. Pew Research Center: Internet, Science & Tech. Published May 12, 2011. Accessed October 10, 2021. https://www.pewresearch.org/internet/2011/05/12/the-social-life-of-health-information-2011/
25. Park A, Conway M. Tracking Health Related Discussions on Reddit for Public Health Applications. *AMIA Annu Symp Proc AMIA Symp*. 2017;2017:1362-1371.
26. Sewalk KC, Tuli G, Hswen Y, Brownstein JS, Hawkins JB. Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study. *J Med Internet Res*. 2018;20(10):e10043. doi:10.2196/10043
27. Adrover C, Bodnar T, Huang Z, Telenti A, Salathé M. Identifying Adverse Effects of HIV Drug Treatment and Associated Sentiments Using Twitter. *JMIR Public Health Surveill*. 2015;1(2):e4488. doi:10.2196/publichealth.4488
28. Bour C, Ahne A, Schmitz S, Perchoux C, Dessenne C, Fagherazzi G. The Use of Social Media for Health Research Purposes: Scoping Review. *J Med Internet Res*. 2021;23(5):e25736. doi:10.2196/25736
29. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the Importance of Digital Epidemiology. *N Engl J Med*. 2013;369(5):401. doi:10.1056/NEJMp1307752
30. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev Med*. 2014;63:112-115. doi:10.1016/j.ypmed.2014.01.024
31. Colditz JB, Chu KH, Emery SL, et al. Toward Real-Time Infoveillance of Twitter Health Messages. *Am J Public Health*. 2018;108(8):1009-1014. doi:10.2105/AJPH.2018.304497
32. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res*. 2013;15(4):e85. doi:10.2196/jmir.1933
33. Grindrod K, Forgione A, Tsuyuki RT, Gavura S, Giustini D. Pharmacy 2.0: a scoping review of social media use in pharmacy. *Res Soc Adm Pharm RSAP*. 2014;10(1):256-270. doi:10.1016/j.sapharm.2013.05.004
34. Greysen SR, Kind T, Chretien KC. Online professionalism and the mirror of social media. *J Gen Intern Med*. 2010;25(11):1227-1229. doi:10.1007/s11606-010-1447-1
35. Denecke K, Bamidis P, Bond C, et al. Ethical Issues of Social Media Usage in Healthcare. *Yearb Med Inform*. 2015;10(1):137-147. doi:10.15265/IY-2015-001
36. Liu Y, Mei Q, Hanauer DA, Zheng K, Lee JM. Use of Social Media in the Diabetes Community: An Exploratory Analysis of Diabetes-Related Tweets. *JMIR Diabetes*. 2016;1(2):e6256. doi:10.2196/diabetes.6256

37. Tenderich A, Tenderich B, Barton T, Richards SE. What Are PWDs (People With Diabetes) Doing Online? A Netnographic Analysis. *J Diabetes Sci Technol*. 2019;13(2):187-197. doi:10.1177/1932296818813192
38. Litchman ML, Walker HR, Ng AH, et al. State of the Science: A Scoping Review and Gap Analysis of Diabetes Online Communities. *J Diabetes Sci Technol*. 2019;13(3):466-492. doi:10.1177/1932296819831042
39. Miller TA, Dimatteo MR. Importance of family/social support and impact on adherence to diabetic therapy. *Diabetes Metab Syndr Obes Targets Ther*. 2013;6:421-426. doi:10.2147/DMSO.S36368
40. Beguerisse-Díaz M, McLennan AK, Garduño-Hernández G, Barahona M, Ulijaszek SJ. The 'who' and 'what' of #diabetes on Twitter. *Digit Health*. 2017;3:2055207616688841. doi:10.1177/2055207616688841
41. Toma T, Athanasiou T, Harling L, Darzi A, Ashrafian H. Online social networking services in the management of patients with diabetes mellitus: systematic review and meta-analysis of randomised controlled trials. *Diabetes Res Clin Pract*. 2014;106(2):200-211. doi:10.1016/j.diabres.2014.06.008
42. Alanzi T. Role of Social Media in Diabetes Management in the Middle East Region: Systematic Review. *J Med Internet Res*. 2018;20(2):e9190. doi:10.2196/jmir.9190
43. Petrovski G, Zivkovic M. Are We Ready to Treat Our Diabetes Patients Using Social Media? Yes, We Are. *J Diabetes Sci Technol*. 2019;13(2):171-175. doi:10.1177/1932296818795441
44. Nelakurthi AR, Pinto AM, Cook CB, et al. Should patients with diabetes be encouraged to integrate social media into their care plan? https://doi.org/10.4155/fsoa-2018-0021. doi:10.4155/fsoa-2018-0021
45. Mandl KD, McNabb M, Marks N, et al. Participatory surveillance of diabetes device safety: a social media-based complement to traditional FDA reporting. *J Am Med Inform Assoc*. 2014;21(4):687-691. doi:10.1136/amiajnl-2013-002127
46. Weitzman ER, Kelemen S, Quinn M, Eggleston EM, Mandl KD. Participatory Surveillance of Hypoglycemia and Harms in an Online Social Network. *JAMA Intern Med*. 2013;173(5):345-351. doi:10.1001/jamainternmed.2013.2512
47. Kubben P, Dumontier M, Dekker A, eds. *Fundamentals of Clinical Data Science*. Springer International Publishing; 2019. doi:10.1007/978-3-319-99713-1
48. Mamlin BW, Tierney WM. The Promise of Information and Communication Technology in Healthcare: Extracting Value From the Chaos. *Am J Med Sci*. 2016;351(1):59-68. doi:10.1016/j.amjms.2015.10.015
49. Coiera E. *Guide to Health Informatics*. 2nd ed. CRC Press; 2003. doi:10.1201/b13617
50. Fieschi M, Dufour JC, Staccini P, Gouvernet J, Bouhaddou O. Medical decision support systems: old dilemmas and new paradigms? *Methods Inf Med*. 2003;42(3):190-198.
51. Khairat S, Marc D, Crosby W, Sanousi AA. Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR Med Inform*. 2018;6(2):e8912. doi:10.2196/medinform.8912
52. Bates DW, Kuperman GJ, Wang S, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc JAMIA*. 2003;10(6):523-530. doi:10.1197/jamia.M1370
53. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. *JAMA*. 2017;318(6):517-518. doi:10.1001/jama.2017.7797
54. Sanchez-Martinez S, Camara O, Piella G, et al. Machine Learning for Clinical Decision-Making: Challenges and Opportunities. *Prepr 2019*. Published online November 24, 2019. doi:10.20944/preprints201911.0278.v1

55. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. *J Am Med Inform Assoc JAMIA*. 2007;14(2):141-145. doi:10.1197/jamia.M2334
56. Rampil IJ. The National Library of Medicine PubMed. *Anesthesiology*. 1997;87(5):1268-1268. doi:10.1097/00000542-199711000-00053
57. El Naqa I, Murphy MJ. What Is Machine Learning? In: El Naqa I, Li R, Murphy MJ, eds. *Machine Learning in Radiation Oncology: Theory and Applications*. Springer International Publishing; 2015:3-11. doi:10.1007/978-3-319-18305-3_1
58. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*. 1943;5(4):115-133. doi:10.1007/BF02478259
59. O'Shea K, Nash R. An Introduction to Convolutional Neural Networks. *ArXiv151108458 Cs*. Published online December 2, 2015. Accessed October 1, 2021. http://arxiv.org/abs/1511.08458
60. Chowdhury GG. Natural language processing. *Annu Rev Inf Sci Technol*. 2003;37(1):51-89. doi:10.1002/aris.1440370103
61. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. *ArXiv13104546 Cs Stat*. Published online October 16, 2013. Accessed November 13, 2020. http://arxiv.org/abs/1310.4546
62. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. *ArXiv13013781 Cs*. Published online September 6, 2013. Accessed November 13, 2020. http://arxiv.org/abs/1301.3781
63. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. *ArXiv160704606 Cs*. Published online June 19, 2017. Accessed November 13, 2020. http://arxiv.org/abs/1607.04606
64. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs*. Published online May 24, 2019. Accessed November 13, 2020. http://arxiv.org/abs/1810.04805
65. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *ArXiv170603762 Cs*. Published online December 5, 2017. Accessed June 17, 2021. http://arxiv.org/abs/1706.03762
66. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018
67. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: *COLT '92*. ; 1992. doi:10.1145/130385.130401
68. Na S, Xumin L, Yong G. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. *2010 Third Int Symp Intell Inf Technol Secur Inform*. Published online 2010. doi:10.1109/IITSI.2010.74
69. Zhao Y, Karypis G. Evaluation of hierarchical clustering algorithms for document datasets. In: *CIKM '02*. ; 2002. doi:10.1145/584792.584877
70. Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ; 2001:282-289.
71. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. *J Biomed Inform*. 2015;58:11-18. doi:10.1016/j.jbi.2015.09.010
72. Settles B. Active Learning Literature Survey. Published online 2010. Accessed December 19, 2020. http://burrsettles.com/pub/settles.activelearning.pdf?source=post_page
73. Lewis DD, Gale WA. A sequential algorithm for training text classifiers. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information*

Retrieval. SIGIR '94. Springer-Verlag; 1994:3-12.
74. Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth Int Conf Weblogs Soc Media ICWSM-14*. Published online 2014:10.
75. U.S. Census Bureau. 2013-2017 American community survey 5-year estimates using American FactFinder. Published December 8, 2019. Accessed August 12, 2019. http://factfinder.census.gov/
76. Conner F, Pfiester E, Elliott J, Slama-Chaudhry A. Unaffordable insulin: patients pay the price. *Lancet Diabetes Endocrinol*. 2019;7(10):748. doi:10.1016/S2213-8587(19)30260-8
77. Fralick M, Kesselheim A. The U.S. Insulin Crisis - Rationing a Lifesaving Medication Discovered in the 1920s. *N Engl J Med*. Published online 2019. doi:10.1056/NEJMp1909402
78. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets. *ArXiv200510200 Cs*. Published online October 5, 2020. Accessed June 17, 2021. http://arxiv.org/abs/2005.10200
79. Ahne A, Orchard F, Tannier X, et al. Insulin pricing and other major diabetes-related concerns in the USA: a study of 46 407 tweets between 2017 and 2019. *BMJ Open Diabetes Res Care*. 2020;8(1):e001190. doi:10.1136/bmjdrc-2020-001190
80. Doan S, Yang EW, Tilak S, Li P, Zisook D, Torii M. Extracting health-related causality from twitter messages using natural language processing. *BMC Med Inform Decis Mak*. 2019;19(Suppl 3):79. doi:10.1186/s12911-019-0785-0
81. Tannier X. NLP-driven Data Journalism: Time-Aware Mining and Visualization of International Alliances. In: *Natural Language Meets Journalism*. ; 2016:5.
82. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2018;46(D1):D8-D13. doi:10.1093/nar/gkx1095
83. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.
84. Zhang Y, Lease M, Wallace BC. Active Discriminative Text Representation Learning. *ArXiv160604212 Cs*. Published online December 1, 2016. Accessed November 23, 2020. http://arxiv.org/abs/1606.04212
85. Vellido Alcacena A, Martín JD, Rossi F, Lisboa PJG. Seeing is believing: the importance of visualization in real-world machine learning applications. In: ; 2011:219-226. Accessed December 9, 2020. https://upcommons.upc.edu/handle/2117/20273
86. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*. 2018;320(21):2199-2200. doi:10.1001/jama.2018.17163
87. Lage I, Ross AS, Kim B, Gershman SJ, Doshi-Velez F. Human-in-the-Loop Interpretability Prior. *ArXiv180511571 Cs Stat*. Published online October 30, 2018. Accessed January 12, 2021. http://arxiv.org/abs/1805.11571
88. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns*. 2021;2(10). doi:10.1016/j.patter.2021.100347
89. Merrill RM. *Introduction to Epidemiology*. Jones & Bartlett Learning; 2010.
90. Valleron AJ. Brève histoire de l'épidémiologie avant le XXe siècle. *Epidémiologie*. Published online 2011.
91. Rohrbasser JM. John Graunt et les bulletins de Londres : une statistique de la mortalité au XVIIe siècle. *Dix-Septieme Siecle*. 2009;n° 243(2):345-368.
92. Morabia A. P. C. A. Louis and the birth of clinical epidemiology. *J Clin Epidemiol*. 1996;49(12):1327-1333. doi:10.1016/s0895-4356(96)00294-6
93. Centers for Disease Control and Prevention (CDC). Introduction to Epidemiology - Section 2: Historical Evolution of Epidemiology. Published 2021. Accessed July 10, 2021. https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section2.html

94. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *Br Med J*. 1950;2(4682):739-748. doi:10.1136/bmj.2.4682.739
95. Fenner F, Henderson D, Arita I, Jezek Z, Ladnyi I. *Smallpox and Its Eradication*. World Health Organization - Geneva; 1988.
96. Müller MM. On the use of applied machine learning and digital infrastructure to leverage social media data in health and epidemiology. Published online 2021. 10.5075/epfl-thesis-8283
97. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res*. 2009;11(1):e11. doi:10.2196/jmir.1157
98. Salathé M. Digital epidemiology: what is it, and where is it going? *Life Sci Soc Policy*. 2018;14:1. doi:10.1186/s40504-017-0065-7
99. Guillory J, Wiant KF, Farrelly M, et al. Recruiting Hard-to-Reach Populations for Survey Research: Using Facebook and Instagram Advertisements and In-Person Intercept in LGBT Bars and Nightclubs to Recruit LGBT Young Adults. *J Med Internet Res*. 2018;20(6):e197. doi:10.2196/jmir.9461
100. Kayrouz R, Dear BF, Karin E, Titov N. Facebook as an effective recruitment strategy for mental health research of hard to reach populations. *Internet Interv*. 2016;4:1-10. doi:10.1016/j.invent.2016.01.001
101. Klingwort J, Schnell R. Critical Limitations of Digital Epidemiology: *Surv Res Methods*. 2020;14(2):95-101. doi:10.18148/srm/2020.v14i2.7726
102. Caplan A, Friesen P. Health disparities and clinical trial recruitment: Is there a duty to tweet? *PLOS Biol*. 2017;15(3):e2002040. doi:10.1371/journal.pbio.2002040
103. Khatri C, Chapman SJ, Glasbey J, et al. Social Media and Internet Driven Study Recruitment: Evaluating a New Model for Promoting Collaborator Engagement and Participation. Sullivan PS, ed. *PLOS ONE*. 2015;10(3):e0118899. doi:10.1371/journal.pone.0118899
104. Thornton L, Batterham PJ, Fassnacht DB, Kay-Lambkin F, Calear AL, Hunt S. Recruiting for health, medical or psychosocial research using Facebook: Systematic review. *Internet Interv*. 2016;4:72-81. doi:10.1016/j.invent.2016.02.001
105. Sanchez C, Grzenda A, Varias A, et al. Social media recruitment for mental health research: A systematic review. *Compr Psychiatry*. 2020;103:152197. doi:10.1016/j.comppsych.2020.152197
106. Fagherazzi G, Ravaud P. Digital diabetes: Perspectives for diabetes prevention, management and research. *Diabetes Metab*. 2019;45(4):322-329. doi:10.1016/j.diabet.2018.08.012
107. Hou C, Carter B, Hewitt J, Francisa T, Mayor S. Do Mobile Phone Applications Improve Glycemic Control (HbA1c) in the Self-management of Diabetes? A Systematic Review, Meta-analysis, and GRADE of 14 Randomized Trials. *Diabetes Care*. 2016;39(11):2089-2095. doi:10.2337/dc16-0346
108. Zeevi D, Korem T, Zmora N, et al. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*. 2015;163(5):1079-1094. doi:10.1016/j.cell.2015.11.001
109. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6
110. Feig DS, Donovan LE, Corcoy R, et al. Continuous glucose monitoring in pregnant women with type 1 diabetes (CONCEPTT): a multicentre international randomised controlled trial. *Lancet Lond Engl*. 2017;390(10110):2347-2359. doi:10.1016/S0140-6736(17)32400-5
111. Dunn TC, Xu Y, Hayter G, Ajjan RA. Real-world flash glucose monitoring patterns and

associations between self-monitoring frequency and glycaemic measures: A European analysis of over 60 million glucose tests. *Diabetes Res Clin Pract*. 2018;137:37-46. doi:10.1016/j.diabres.2017.12.015
112. Quemerais MA, Doron M, Dutrech F, et al. Preliminary evaluation of a new semi-closed-loop insulin therapy system over the prandial period in adult patients with type 1 diabetes: the WP6.0 Diabeloop study. *J Diabetes Sci Technol*. 2014;8(6):1177-1184. doi:10.1177/1932296814545668
113. Bates M. Tracking Disease: Digital Epidemiology Offers New Promise in Predicting Outbreaks. *IEEE Pulse*. 2017;8(1):18-22. doi:10.1109/MPUL.2016.2627238
114. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056
115. Kiranyaz S, Ince T, Gabbouj M. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Trans Biomed Eng*. 2016;63(3):664-675. doi:10.1109/TBME.2015.2468589
116. Fagherazzi G. Deep Digital Phenotyping and Digital Twins for Precision Health: Time to Dig Deeper. *J Med Internet Res*. 2020;22(3):e16770. doi:10.2196/16770
117. Mittelstadt B, Benzler J, Engelmann L, Prainsack B, Vayena E. Is there a duty to participate in digital epidemiology? *Life Sci Soc Policy*. 2018;14(1):9. doi:10.1186/s40504-018-0074-1
118. Gerke S, Minssen T, Cohen G. Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, eds. *Artificial Intelligence in Healthcare*. Academic Press; 2020:295-336. doi:10.1016/B978-0-12-818438-7.00012-5
119. US Department of Homeland Security. Cybersecurity. Published 2019. Accessed October 27, 2021. https://www.dhs.gov/topic/cybersecurity
120. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*. 2020;323(4):305-306. doi:10.1001/jama.2019.20866
121. Briganti G, Le Moine O. Artificial Intelligence in Medicine: Today and Tomorrow. *Front Med*. 2020;7:27. doi:10.3389/fmed.2020.00027
122. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3(1):9. doi:10.1186/s40537-016-0043-6
123. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal*. 2019;54:280-296. doi:10.1016/j.media.2019.03.009
124. Ebbehoj A, Thunbo M, Andersen OE, Glindtvad MV, Hulman A. *Transfer Learning for Non-Image Data in Clinical Research: A Scoping Review*.; 2021:2021.10.01.21264290. doi:10.1101/2021.10.01.21264290
125. World Health Organization. Fact sheet: Diabetes. Published 2021. Accessed October 11, 2021. https://www.who.int/news-room/fact-sheets/detail/diabetes
126. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*. 2004;27(suppl 1):s5-s10. doi:10.2337/diacare.27.2007.S5
127. Wilcox G. Insulin and insulin resistance. *Clin Biochem Rev*. 2005;26(2):19-39.
128. Selvin E, Parrinello CM, Daya N, Bergenstal RM. Trends in Insulin Use and Diabetes Control in the U.S.: 1988–1994 and 1999–2012. *Diabetes Care*. 2016;39(3):e33-e35. doi:10.2337/dc15-2229
129. Basu S, Yudkin JS, Kehlenbrink S, et al. Estimation of global insulin use for type 2 diabetes, 2018-30: a microsimulation analysis. *Lancet Diabetes Endocrinol*. 2019;7(1):25-33. doi:10.1016/S2213-8587(18)30303-6
130. British Diabetic Association, Diabetes UK. Insulin and diabetes. Published 2021. Accessed October 15, 2021. https://www.diabetes.org.uk/guide-to-diabetes/managing-your-

diabetes/treating-your-diabetes/insulin
131. Roche EF, Menon A, Gill D, Hoey H. Clinical presentation of type 1 diabetes. *Pediatr Diabetes*. 2005;6(2):75-78. doi:10.1111/j.1399-543X.2005.00110.x
132. American Diabetes Association. 3. Prevention or Delay of Type 2 Diabetes: Standards of Medical Care in Diabetes-2021. *Diabetes Care*. 2021;44(Suppl 1):S34-S39. doi:10.2337/dc21-S003
133. Ahmad LA, Crandall JP. Type 2 Diabetes Prevention: A Review. *Clin Diabetes*. 2010;28(2):53-59. doi:10.2337/diaclin.28.2.53
134. Bansal N. Prediabetes diagnosis and treatment: A review. *World J Diabetes*. 2015;6(2):296-303. doi:10.4239/wjd.v6.i2.296
135. Plows JF, Stanley JL, Baker PN, Reynolds CM, Vickers MH. The Pathophysiology of Gestational Diabetes Mellitus. *Int J Mol Sci*. 2018;19(11):3342. doi:10.3390/ijms19113342
136. Finer S, Robb P, Cowan K, Daly A, Robertson E, Farmer A. Top ten research priorities for type 2 diabetes: results from the Diabetes UK–James Lind Alliance Priority Setting Partnership. *Lancet Diabetes Endocrinol*. 2017;5(12):935-936. doi:10.1016/S2213-8587(17)30324-8
137. Fisher L, Gonzalez JS, Polonsky WH. The confusing tale of depression and distress in patients with diabetes: a call for greater clarity and precision. *Diabet Med J Br Diabet Assoc*. 2014;31(7):764-772. doi:10.1111/dme.12428
138. Kiriella DA, Islam S, Oridota O, et al. Unraveling the concepts of distress, burnout, and depression in type 1 diabetes: A scoping review. *EClinicalMedicine*. 2021;40:101118. doi:10.1016/j.eclinm.2021.101118
139. Bruno BA, Choi D, Thorpe KE, Yu CH. Relationship Among Diabetes Distress, Decisional Conflict, Quality of Life, and Patient Perception of Chronic Illness Care in a Cohort of Patients With Type 2 Diabetes and Other Comorbidities. *Diabetes Care*. 2019;42(7):1170-1177. doi:10.2337/dc18-1256
140. Ogbera A, Adeyemi-Doro A. Emotional distress is associated with poor self care in type 2 diabetes mellitus. *J Diabetes*. 2011;3(4):348-352. doi:10.1111/j.1753-0407.2011.00156.x
141. Delahanty LM, Grant RW, Wittenberg E, et al. Association of diabetes-related emotional distress with diabetes treatment in primary care patients with Type 2 diabetes. *Diabet Med J Br Diabet Assoc*. 2007;24(1):48-54. doi:10.1111/j.1464-5491.2007.02028.x
142. Hessler D, Fisher L, Glasgow RE, et al. Reductions in Regimen Distress Are Associated With Improved Management and Glycemic Control Over Time. *Diabetes Care*. 2014;37(3):617-624. doi:10.2337/dc13-0762
143. Schmidt CB, van Loon BJP, Vergouwen ACM, Snoek FJ, Honig A. Systematic review and meta-analysis of psychological interventions in people with diabetes and elevated diabetes-distress. *Diabet Med*. 2018;35(9):1157-1172. doi:10.1111/dme.13709
144. Gundecha P, Liu H. Mining Social Media: A Brief Introduction. In: ; 2012. doi:10.1287/educ.1120.0105
145. Statista. Leading countries based on number of Twitter users as of July 2021. Published 2021. Accessed October 11, 2021. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
146. Litchman ML, Edelman LS, Donaldson GW. Effect of Diabetes Online Community Engagement on Health Indicators: Cross-Sectional Study. *JMIR Diabetes*. 2018;3(2):e8. doi:10.2196/diabetes.8603
147. Antheunis ML, Tates K, Nieboer TE. Patients' and health professionals' use of social media in health care: motives, barriers and expectations. *Patient Educ Couns*. 2013;92(3):426-431. doi:10.1016/j.pec.2013.06.020
148. Gomez-Galvez P, Mejías CS, Fernandez-Luque L. Social media for empowering people with

diabetes: Current status and future trends. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. ; 2015:2135-2138. doi:10.1109/EMBC.2015.7318811
149. Sewalk KC, Tuli G, Hswen Y, Brownstein JS, Hawkins JB. Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study. *J Med Internet Res*. 2018;20(10):e10043. doi:10.2196/10043
150. Davies B, Kotter M. Lessons From Recruitment to an Internet-Based Survey for Degenerative Cervical Myelopathy: Comparison of Free and Fee-Based Methods. *JMIR Res Protoc*. 2018;7(2):e6567. doi:10.2196/resprot.6567
151. Ventola CL. Social Media and Health Care Professionals: Benefits, Risks, and Best Practices. *Pharm Ther*. 2014;39(7):491-520.
152. Chauhan B, George R, Coffin J. Social media and you: what every physician needs to know. *J Med Pract Manag MPM*. 2012;28(3):206-209.
153. Collins K, Shiffman D, Rock J. How Are Scientists Using Social Media in the Workplace? *PLOS ONE*. 2016;11(10):e0162680. doi:10.1371/journal.pone.0162680
154. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. *Health Inf Libr J*. 2017;34(4):268-283. doi:10.1111/hir.12192
155. Holtz B, Smock A, Reyes-Gastelum D. Connected Motherhood: Social Support for Moms and Moms-to-Be on Facebook. *Telemed J E Health*. 2015;21(5):415-421. doi:10.1089/tmj.2014.0118
156. Rozenblum R, Bates DW. Patient-centred healthcare, social media and the internet: the perfect storm? *BMJ Qual Saf*. 2013;22(3):183-186. doi:10.1136/bmjqs-2012-001744
157. Alanzi T, Al-Yami S. Physicians' Attitude towards The Use of Social Media for Professional Purposes in Saudi Arabia. *Int J Telemed Appl*. 2019;2019:6323962. doi:10.1155/2019/6323962
158. Chretien KC, Kind T. Social media and clinical care: ethical, professional, and social implications. *Circulation*. 2013;127(13):1413-1421. doi:10.1161/CIRCULATIONAHA.112.128017
159. Omer T. Empowered citizen 'health hackers' who are not waiting. *BMC Med*. 2016;14(1):118. doi:10.1186/s12916-016-0670-y
160. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. *J Mach Learn Res*. 2003;3:993-1022.
161. Stellefson M, Chaney B, Barry AE, et al. Web 2.0 chronic disease self-management for older adults: a systematic review. *J Med Internet Res*. 2013;15(2):e35. doi:10.2196/jmir.2439
162. Petrovski G, Zivkovic M. Impact of Facebook on Glucose Control in Type 1 Diabetes: A Three-Year Cohort Study. *JMIR Diabetes*. 2017;2(1):e9. doi:10.2196/diabetes.7693
163. Hanberger L, Ludvigsson J, Nordfeldt S. Use of a Web 2.0 Portal to Improve Education and Communication in Young Patients With Families: Randomized Controlled Trial. *J Med Internet Res*. 2013;15(8):e2425. doi:10.2196/jmir.2425
164. Whittemore R, Liberti LS, Jeon S, et al. Efficacy and implementation of an Internet psychoeducational program for teens with type 1 diabetes. *Pediatr Diabetes*. 2016;17(8):567-575. doi:10.1111/pedi.12338
165. Nasar Z, Jaffry SW, Malik MK. Information extraction from scientific articles: a survey. *Scientometrics*. 2018;117(3):1931-1990. doi:10.1007/s11192-018-2921-5
166. Yeo-Teh NSL, Tang BL. An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19). *Account Res*. 2021;28(1):47-53. doi:10.1080/08989621.2020.1782203
167. Ware M, Mabe M. The STM Report: An overview of scientific and scholarly journal publishing. Published online 2015:181.

168. Allahyari M, Pouriyeh S, Assefi M, et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *ArXiv170702919 Cs*. Published online July 28, 2017. Accessed September 27, 2021. http://arxiv.org/abs/1707.02919
169. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312(7023):71-72. doi:10.1136/bmj.312.7023.71
170. Price WN. Big Data and Black-Box Medical Algorithms. *Sci Transl Med*. 2018;10(471):eaao5333. doi:10.1126/scitranslmed.aao5333
171. Simon SR, McCarthy ML, Kaushal R, et al. Electronic health records: which practices have them, and how are clinicians using them? *J Eval Clin Pract*. 2008;14(1):43-47. doi:crf
172. Baeza-Yates R, Ribeiro-Neto B. *Modern Information Retrieval*. ACM Press, New York; 1999.
173. Schmidhuber J. Deep Learning in Neural Networks: An Overview. *Neural Netw*. 2014;61:85-117. doi:10.1016/j.neunet.2014.09.003
174. Bengio Y, Ducharme R, Vincent P, Jauvin C. A Neural Probabilistic Language Model. *J Mach Learn Res*. Published online 2003:19.
175. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM*. 1975;18(11):613-620.
176. Turney PD, Pantel P. From Frequency to Meaning: Vector Space Models of Semantics. *J Artif Intell Res*. 2010;37:141-188. doi:10.1613/jair.2934
177. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014:1532-1543. doi:10.3115/v1/D14-1162
178. Mikolov T, Yih W tau, Zweig G. Linguistic Regularities in Continuous Space Word Representations. In: *HLT-NAACL*. ; 2013.
179. Pal S. Implementing Word2Vec in Tensorflow. Published 2019. Accessed July 30, 2021. https://medium.com/analytics-vidhya/implementing-word2vec-in-tensorflow-44f93cf2665f
180. Alammar J. The illustrated Word2vec. Published 2018. Accessed July 30, 2021. http://jalammar.github.io/illustrated-word2vec/
181. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. JMLR.org; 2014:II-1188-II-1196.
182. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. *ArXiv180205365 Cs*. Published online March 22, 2018. Accessed July 31, 2021. http://arxiv.org/abs/1802.05365
183. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. *ArXiv180106146 Cs Stat*. Published online May 23, 2018. Accessed June 17, 2021. http://arxiv.org/abs/1801.06146
184. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. Published online 2018:12.
185. Schmidt RM. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. *ArXiv191205911 Cs Stat*. Published online November 23, 2019. Accessed August 9, 2021. http://arxiv.org/abs/1912.05911
186. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735
187. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *Proc 3rd Int Conf Learn Represent*. Published online 2015. Accessed August 9, 2021. http://arxiv.org/abs/1409.0473

188. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv190711692 Cs*. Published online July 26, 2019. Accessed June 17, 2021. http://arxiv.org/abs/1907.11692
189. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv191001108 Cs*. Published online February 29, 2020. Accessed August 10, 2021. http://arxiv.org/abs/1910.01108
190. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published online September 10, 2019:btz682. doi:10.1093/bioinformatics/btz682
191. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018
192. Duda R, Hart P, G.Stork D. Pattern Classification. In: *Wiley Interscience*. ; 2001.
193. Roux M. A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. *J Classif*. 2018;35(2):345-366. doi:10.1007/s00357-018-9259-9
194. Sutton C. An Introduction to Conditional Random Fields. *Found Trends® Mach Learn*. 2012;4(4):267-373. doi:10.1561/2200000013
195. Rocktäschel T, Huber T, Weidlich M, Leser U. WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In: ; 2013.
196. Konkol M, Konopík M. CRF-Based Czech Named Entity Recognizer and Consolidation of Czech NER Research. In: Habernal I, Matoušek V, eds. *Text, Speech, and Dialogue*. Lecture Notes in Computer Science. Springer; 2013:153-160. doi:10.1007/978-3-642-40585-3_20
197. Şeker GA, Eryiğit G. Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content1. Hotho A, Jäschke R, Lerman K, Hotho A, Jäschke R, Lerman K, eds. *Semantic Web*. 2017;8(5):625-642. doi:10.3233/SW-170253
198. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*. 2017;33(14):i37-i48. doi:10.1093/bioinformatics/btx228
199. Passos A, Kumar V, McCallum A. Lexicon Infused Phrase Embeddings for Named Entity Resolution. *ArXiv14045367 Cs*. Published online April 21, 2014. Accessed August 12, 2021. http://arxiv.org/abs/1404.5367
200. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2016:260-270. doi:10.18653/v1/N16-1030
201. Ramshaw LA, Marcus MP. Text Chunking using Transformation-Based Learning. *ArXivcmp-Lg9505040*. Published online May 23, 1995. Accessed June 21, 2021. http://arxiv.org/abs/cmp-lg/9505040
202. Yang Y, Katiyar A. Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning. *ArXiv201002405 Cs*. Published online October 5, 2020. Accessed October 11, 2021. http://arxiv.org/abs/2010.02405
203. Seung HS, Opper M, Sompolinsky H. Query by committee. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Association for Computing Machinery; 1992:287-294. doi:10.1145/130385.130417
204. Settles B, Craven M, Ray S. Multiple-instance active learning. In: *MIT Press*. Vol 20. ; 2008:1289-1296.
205. Tong S, Koller D. Support vector machine active learning with applications to text classification. *J Mach Learn Res*. 2002;2:45-66. doi:10.1162/153244302760185243
206. McCallum AK, Nigam K. Employing EM and Pool-Based Active Learning for Text

Classification. Published online 1998:9.
207. Lu J, Henchion M, Mac Namee B. Investigating the Effectiveness of Representations Based on Word-Embeddings in Active Learning for Labelling Text Datasets. *ArXiv191003505 Cs Stat*. Published online October 10, 2019. Accessed November 23, 2020. http://arxiv.org/abs/1910.03505
208. Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. *ArXiv200805756 Cs Stat*. Published online August 13, 2020. Accessed August 15, 2021. http://arxiv.org/abs/2008.05756
209. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7
210. Larsen B, Aone C. Fast and effective text mining using linear-time document clustering. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '99. Association for Computing Machinery; 1999:16-22. doi:10.1145/312129.312186
211. Van Rossum G, Drake FL. *Python 3 Reference Manual*. CreateSpace; 2009.
212. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neural Inf Process Syst*. 2019;32:12.
213. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*. ELRA; 2010:45-50.
214. Bostock M, Ogievetsky V, Heer J. D^3 Data-Driven Documents. *IEEE Trans Vis Comput Graph*. 2011;17(12):2301-2309. doi:10.1109/TVCG.2011.185
215. Odersky M, Spoon L, Venners B. *Programming in Scala*. Artima Inc; 2008. https://www.scala-lang.org/
216. Zaharia M, Xin RS, Wendell P, et al. Apache Spark: a unified engine for big data processing. *Commun ACM*. 2016;59(11):56-65. doi:10.1145/2934664
217. Apache Software Foundation. *Apache Lucene*.; 2019. https://lucene.apache.org
218. Schmitt A, Reimer A, Kulzer B, Haak T, Ehrmann D, Hermanns N. How to assess diabetes distress: comparison of the Problem Areas in Diabetes Scale (PAID) and the Diabetes Distress Scale (DDS). *Diabet Med*. 2016;33(6):835-843. doi:10.1111/dme.12887
219. Patel KD, Heppner A, Srivastava G, Mago V. Analyzing use of Twitter by diabetes online community. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '19. Association for Computing Machinery; 2019:937-944. doi:10.1145/3341161.3343673
220. Khatua A, Khatua A, Cambria E. A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Inf Process Manag*. Published online 2019. doi:10.1016/J.IPM.2018.10.010
221. Xu A, Liu Z, Guo Y, Sinha V, Akkiraju R. A New Chatbot for Customer Service on Social Media. In: ; 2017. doi:10.1145/3025453.3025496
222. Johnsen JA, Eggesvik T, Rørvik T, Hanssen M, Wynn R, Kummervold P. Differences in Emotional and Pain-Related Language in Tweets About Dentists and Medical Doctors: Text Analysis of Twitter Content. *JMIR Public Health Surveill*. 2019;5:e10432. doi:10.2196/publichealth.10432
223. Zhang R, Liu N. Recognizing Humor on Twitter. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. CIKM '14. Association for Computing Machinery; 2014:889-898. doi:10.1145/2661829.2661997
224. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953

225. Lemaıtre G, Nogueira F. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J Mach Learn Res*. Published online 2017:5.
226. Hecht B, Hong L, Suh B, Chi EH. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery; 2011:237-246. Accessed August 23, 2021. https://doi.org/10.1145/1978942.1978976
227. Cheng Z, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*. ACM Press; 2010:759. doi:10.1145/1871437.1871535
228. Miura Y, Taniguchi M, Taniguchi T, Ohkuma T. A Simple Scalable Neural Networks based Model for Geolocation Prediction in Twitter. *Proc 2nd Workshop Noisy WNUT*. Published online 2016:235-239.
229. Shah Z, Martin P, Coiera E, Mandl KD, Dunn AG. Modeling Spatiotemporal Factors Associated With Sentiment on Twitter: Synthesis and Suggestions for Improving the Identification of Localized Deviations. *J Med Internet Res*. 2019;21(5):e12881. doi:10.2196/12881
230. GeoNames geographical Database. Published August 20, 2021. https://www.geonames.org/
231. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. *Sov Phys Dokl*. Published online 1965. Accessed August 30, 2021. https://www.semanticscholar.org/paper/Binary-codes-capable-of-correcting-deletions%2C-and-Levenshtein/b2f8876482c97e804bb50a5e2433881ae31d0cdd
232. Lin HT, Lin CJ, Weng RC. A note on Platt's probabilistic outputs for support vector machines. *Mach Learn*. 2007;68(3):267-276. doi:10.1007/s10994-007-5018-6
233. Parrott WG. *Emotions in Social Psychology: Essential Readings*. Psychology Press; 2001.
234. Miller GA. WordNet: A Lexical Database for English. In: *Communications of the ACM*. Vol 38. ; 1995:No. 11: 39-41.
235. Princeton University. *About WordNet*. Princeton University; 2010. https://wordnet.princeton.edu/ (accessed: 22/08/2021)
236. Comesaña M, Soares A, Perea M, Piñeiro Barreiro A, Fraga I, Pinheiro A. ERP correlates of masked affective priming with emoticons. *Comput Hum Behav*. 2013;29:588-595. doi:10.1016/j.chb.2012.10.020
237. Wolny W. Emotion Analysis of Twitter Data That Use Emoticons and Emoji Ideograms. In: *Information Systems Development: Complexity in Information Systems Development (ISD2016 Proceedings)*. ; 2016.
238. Gohil S, Vuik S, Darzi A. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. *JMIR Public Health Surveill*. 2018;4(2):e5789. doi:10.2196/publichealth.5789
239. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning. *J Med Internet Res*. 2015;17(8):e4392. doi:10.2196/jmir.4392
240. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets. *JMIR Public Health Surveill*. 2016;2(2):e6327. doi:10.2196/publichealth.6327
241. Hawkins JB, Brownstein JS, Tuli G, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf*. 2016;25(6):404-413. doi:10.1136/bmjqs-2015-004309
242. Adrover C, Bodnar T, Salathe M. Targeting HIV-related Medication Side Effects and Sentiment Using Twitter Data. *ArXiv14043610 Cs*. Published online April 10, 2014. Accessed August 31, 2021. http://arxiv.org/abs/1404.3610
243. Esuli A, Sebastiani F. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion

Mining. *Proc 5th Conf Lang Resour Eval LREC06*. Published online 2006:6.
244. Zimbra D, Abbasi A, Zeng D, Chen H. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans Manag Inf Syst*. 2018;9(2):5:1-5:29. doi:10.1145/3185045
245. Ribeiro FN, Araújo M, Gonçalves P, Gonçalves MA, Benevenuto F. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci*. Published online 2016. doi:10.1140/epjds/s13688-016-0085-1
246. Tamersoy A, De Choudhury M, Chau DH. Characterizing Smoking and Drinking Abstinence from Social Media. *HT Proc ACM Conf Hypertext Soc Media ACM Conf Hypertext Soc Media*. 2015;2015:139-148. doi:10.1145/2700171.2791247
247. US Social Security Administration. Popular Baby Names. Published 2019. Accessed August 12, 2019. https://www.ssa.gov/oact/babynames/limits.html
248. Mckinney W. Data Structures for Statistical Computing in Python. *Proc 9th Python Sci Conf*. Published online January 1, 2010:56-61.
249. Vishal A, Sonawane SS. Sentiment Analysis of Twitter Data: A Survey of Techniques. *Int J Comput Appl*. 2016;139(11):5-15. doi:10.5120/ijca2016908625
250. Naz S, Sharan A, Malik N. Sentiment Classification on Twitter Data Using Support Vector Machine. In: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. ; 2018:676-679. doi:10.1109/WI.2018.00-13
251. U.S. Census Bureau. State Population Totals: 2010-2019. Published 2021. Accessed August 31, 2021. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
252. Vydiswaran VGV, Romero DM, Zhao X, et al. Uncovering the relationship between food-related discussion on Twitter and neighborhood characteristics. *J Am Med Inform Assoc*. 2020;27(2):254-264. doi:10.1093/jamia/ocz181
253. Nguyen QC, Brunisholz KD, Yu W, et al. Twitter-derived neighborhood characteristics associated with obesity and diabetes. *Sci Rep*. 2017;7(1):16425. doi:10.1038/s41598-017-16573-1
254. Hughes AE, Berg CA, Wiebe DJ. Emotional processing and self-control in adolescents with type 1 diabetes. *J Pediatr Psychol*. 2012;37(8):925-934. doi:10.1093/jpepsy/jss062
255. Coccaro EF, Lazarus S, Joseph J, et al. Emotional Regulation and Diabetes Distress in Adults With Type 1 and Type 2 Diabetes. *Diabetes Care*. 2021;44(1):20-25. doi:10.2337/dc20-1059
256. Hagger V, Hendrieckx C, Sturt J, Skinner TC, Speight J. Diabetes Distress Among Adolescents with Type 1 Diabetes: a Systematic Review. *Curr Diab Rep*. 2016;16(1):9. doi:10.1007/s11892-015-0694-2
257. Richman LS, Kubzansky L, Maselko J, Kawachi I, Choo P, Bauer M. Positive emotion and health: going beyond the negative. *Health Psychol Off J Div Health Psychol Am Psychol Assoc*. 2005;24(4):422-429. doi:10.1037/0278-6133.24.4.422
258. Iturralde E, Chi FW, Grant RW, et al. Association of Anxiety With High-Cost Health Care Use Among Individuals With Type 2 Diabetes. *Diabetes Care*. 2019;42(9):1669-1674. doi:10.2337/dc18-1553
259. T1 International. T1 International insulin4all. Published 2021. Accessed January 8, 2021. https://www.t1international.com/insulin4all/
260. Blanchette JE, Toly VB, Wood JR. Financial stress in emerging adults with type 1 diabetes in the United States. *Pediatr Diabetes*. 2021;22(5):807-815. doi:10.1111/pedi.13216
261. Luo J, Gonsalves G, Greene J. Insulin for all: treatment activism and the global diabetes crisis. *The Lancet*. 2019;393(10186):2116-2117. doi:10.1016/S0140-6736(19)31090-6
262. Rajkumar SV. The High Cost of Insulin in the United States: An Urgent Call to Action. *Mayo*

Clin Proc. 2020;95(1):22-28. doi:10.1016/j.mayocp.2019.11.013
263. Statista. Percentage of U.S. adults who use Twitter as of February 2019, by age group. Published October 28, 2019. https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/
264. Mihăilă C, Ananiadou S. Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomed Eng OnLine*. 2014;13:S1. doi:10.1186/1475-925X-13-S2-S1
265. Khoo CSG, Chan S, Niu Y. Extracting causal knowledge from a medical database using graphical patterns. In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*. Association for Computational Linguistics; 2000:336-343. doi:10.3115/1075218.1075261
266. Qiu J, Xu L, Zhai J, Luo L. Extracting Causal Relations from Emergency Cases Based on Conditional Random Fields. *Procedia Comput Sci*. 2017;112:1623-1632. doi:10.1016/j.procs.2017.08.252
267. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *J Am Med Inform Assoc JAMIA*. 2017;24(4):813-821. doi:10.1093/jamia/ocw180
268. Doan S, Ritchart A, Perry N, Chaparro JD, Conway M. How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets. *JMIR Public Health Surveill*. 2017;3(2):e35. doi:10.2196/publichealth.5939
269. Kayesh H, Islam MS, Wang J. On Event Causality Detection in Tweets. *ArXiv190103526 Cs*. Published online January 11, 2019. Accessed January 20, 2021. http://arxiv.org/abs/1901.03526
270. Yang J, Han SC, Poon J. A Survey on Extraction of Causal Relations from Natural Language Text. *ArXiv210106426 Cs*. Published online January 16, 2021. Accessed June 30, 2021. http://arxiv.org/abs/2101.06426
271. Khoo C, Chan S, Niu Y. The Many Facets of the Cause-Effect Relation. In: Green R, Bean CA, Myaeng SH, eds. *The Semantics of Relationships: An Interdisciplinary Perspective*. Information Science and Knowledge Management. Springer Netherlands; 2002:51-70. doi:10.1007/978-94-017-0073-3_4
272. Asghar N. Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey. *ArXiv160507895 Cs*. Published online May 25, 2016. Accessed January 18, 2021. http://arxiv.org/abs/1605.07895
273. Garcia D. COATIS, an NLP system to locate expressions of actions connected by causality links. In: Plaza E, Benjamins R, eds. *Knowledge Acquisition, Modeling and Management*. Lecture Notes in Computer Science. Springer; 1997:347-352. doi:10.1007/BFb0026799
274. Blanco E, Castell N, Moldovan D. Causal Relation Extraction. In: *LREC*. ; 2008.
275. Kim HD, Castellanos M, Hsu M, Zhai C, Rietz T, Diermeier D. Mining causal topics in text data: iterative topic modeling with time series feedback. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. CIKM '13. Association for Computing Machinery; 2013:885-890. doi:10.1145/2505515.2505612
276. Girju R. Automatic detection of causal relations for Question Answering. In: *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering -*. Vol 12. Association for Computational Linguistics; 2003:76-83. doi:10.3115/1119312.1119322
277. Xu Y, Mou L, Li G, Chen Y, Peng H, Jin Z. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths. In: ; 2015:1785-1794. doi:10.18653/v1/D15-1206
278. Wang L, Cao Z, Melo G de, Liu Z. Relation Classification via Multi-Level Attention CNNs. In: ; 2016:1298-1307. doi:10.18653/v1/P16-1123

279. Ponti EM, Korhonen A. Event-Related Features in Feedforward Neural Networks Contribute to Identifying Causal Relations in Discourse. In: ; 2017:25-30. doi:10.18653/v1/W17-0903
280. Khetan V, Ramnani R, Anand M, Sengupta S, Fano AE. Causal BERT: Language Models for Causality Detection Between Events Expressed in Text. In: Arai K, ed. *Intelligent Computing*. Lecture Notes in Networks and Systems. Springer International Publishing; 2021:965-980. doi:10.1007/978-3-030-80119-9_64
281. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Published online 2012. doi:10.1016/j.jbi.2012.04.008
282. Ahne A, Khetan V, Tannier X, et al. Identifying causal associations in tweets using deep learning: Use case on diabetes-related tweets from 2017-2021. *ArXiv211101225 Cs*. Published online November 4, 2021. Accessed November 6, 2021. http://arxiv.org/abs/2111.01225
283. Wolf T, Debut L, Sanh V, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv191003771 Cs*. Published online July 13, 2020. Accessed June 17, 2021. http://arxiv.org/abs/1910.03771
284. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20(1):37-46. doi:10.1177/001316446002000104
285. Altman D. *Practical Statistics for Medical Research*. Chapman & Hall; 1991.
286. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
287. Alshammari N, Alanazi S. The impact of using different annotation schemes on named entity recognition. *Egypt Inform J*. 2021;22(3):295-302. doi:10.1016/j.eij.2020.10.004
288. Korobov M. *Sklearn-Crfsuite: CRFsuite Wrapper*.; 2021. Accessed August 15, 2021. https://sklearn-crfsuite.readthedocs.io/en/latest/index.html
289. Okazaki N. *CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs)*.; 2007. Accessed August 16, 2021. http://www.chokkan.org/software/crfsuite/
290. Bollegala D, Maskell S, Sloane R, Hajne J, Pirmohamed M. Causality Patterns for Detecting Adverse Drug Reactions From Social Media: Text Mining Approach. *JMIR Public Health Surveill*. 2018;4(2):e8214. doi:10.2196/publichealth.8214
291. Dasgupta T, Saha R, Dey L, Naskar A. Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks | SIGdial 2018 - video recordings and slides. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Di- Alogue*. ; 2019:306-316. Accessed October 13, 2021. https://www.superlectures.com/sigdial2018/automatic-extraction-of-causal-relations-from-text-using-linguistically-informed-deep-neural-networks
292. Khetan V, Rizvi MIH, Huber J, Bartusiak P, Sacaleanu B, Fano A. MIMICause : Defining, identifying and predicting types of causal relationships between biomedical concepts from clinical notes. *ArXiv211007090 Cs*. Published online October 13, 2021. Accessed October 15, 2021. http://arxiv.org/abs/2110.07090
293. Statista. Percentage of U.S. adults who use Twitter as of February 2021, by age group. Published 2021. Accessed May 10, 2021. https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/
294. Ahne A, Fagherazzi G, Tannier X, Czernichow T, Orchard F. Improving Diabetes-Related Biomedical Literature Exploration in the Clinical Decision-making Process via Interactive Classification and Topic Discovery: Methodology Development Study. *J Med Internet Res*. 2022;24(1):e27434. doi:10.2196/27434
295. Masic I, Miokovic M, Muhamedagic B. Evidence based medicine - new approaches and challenges. *Acta Inform Medica AIM J Soc Med Inform Bosnia Herzeg Cas Drustva Za Med*

Inform BiH. 2008;16(4):219-225. doi:10.5455/aim.2008.16.219-225
296. Hernandez-Medrano I, Guijarro J, Belda C, et al. Savana. Re-using Electronic Health Records with Artificial Intelligence. *Int J Interact Multimed Artif Intell*. 2017;In Press:1. doi:10.9781/ijimai.2017.03.001
297. van Dijk N, Hooft L, Wieringa-de Waard M. What Are the Barriers to Residents' Practicing Evidence-Based Medicine? A Systematic Review. *Acad Med*. 2010;85(7):1163-1170. doi:10.1097/ACM.0b013e3181d4152f
298. Lyman J, Cohn W, Bloomrosen M, Detmer D. Clinical decision support: progress and opportunities. *J Am Med Inform Assoc*. Published online 2010. doi:10.1136/jamia.2010.005561
299. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. *J Biomed Inform*. 2008;41(2):387-392. doi:10.1016/j.jbi.2007.09.003
300. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet Lond Engl*. 2017;390(10092):415-423. doi:10.1016/S0140-6736(16)31592-6
301. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391
302. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed Inform Insights*. 2016;8:BII.S31559. doi:10.4137/BII.S31559
303. Althoff T, Clark K, Leskovec J. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Trans Assoc Comput Linguist*. 2016;4:463-476.
304. Gkotsis G, Oellrich A, Velupillai S, et al. Characterisation of mental health conditions in social media using Informed Deep Learning. *Sci Rep*. 2017;7:45141. doi:10.1038/srep45141
305. Xu H, Fu Z, Shah A, et al. Extracting and Integrating Data from Entire Electronic Health Records for Detecting Colorectal Cancer Cases. *AMIA Annu Symp Proc*. 2011;2011:1564.
306. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of Machine Learning Techniques for Heart Failure Readmissions. *Circ Cardiovasc Qual Outcomes*. 2016;9(6):629-640. doi:10.1161/CIRCOUTCOMES.116.003039
307. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform*. 2009;42(5):760-772. doi:10.1016/j.jbi.2009.08.007
308. Editorial. Towards trustable machine learning. *Nat Biomed Eng*. 2018;2(10):709-710. doi:10.1038/s41551-018-0315-x
309. Shah NH, Milstein A, Bagley P Steven C. Making Machine Learning Models Clinically Useful. *JAMA*. 2019;322(14):1351-1352. doi:10.1001/jama.2019.10306
310. Abid A, Abdalla A, Abid A, Khan D, Alfozan A, Zou J. An online platform for interactive feedback in biomedical machine learning. *Nat Mach Intell*. 2020;2(2):86-88. doi:10.1038/s42256-020-0147-8
311. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput Appl*. 2020;32(24):18069-18083. doi:10.1007/s00521-019-04051-w
312. Ravì D, Wong C, Deligianni F, et al. Deep Learning for Health Informatics. *IEEE J Biomed Health Inform*. 2017;21(1):4-21. doi:10.1109/JBHI.2016.2636665
313. Bhanot G, Biehl M, Villmann T, Zühlke D. Biomedical data analysis in translational research: Integration of expert knowledge and interpretable models. In: *25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2017*. Ciaco - i6doc.com; 2017:177-186. Accessed December 9, 2020. https://www.rug.nl/research/portal/en/publications/biomedical-data-analysis-in-

translational-research(0d1c2bbf-74ae-4132-9f02-a22e2f6af301).html
314. Lee S, Baker J, Song J, Wetherbe JC. An Empirical Comparison of Four Text Mining Methods. In: *2010 43rd Hawaii International Conference on System Sciences*. IEEE; 2010:1-10. doi:10.1109/HICSS.2010.48
315. Ha-Thuc V, Srinivasan P. Topic models and a revisit of text-related applications. In: *Proceeding of the 2nd PhD Workshop on Information and Knowledge Management - PIKM '08*. ACM Press; 2008:25. doi:10.1145/1458550.1458556
316. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2018;46(D1):D8-D13. doi:10.1093/nar/gkx1095
317. McDonald R, Brokos GI, Androutsopoulos I. Deep Relevance Ranking Using Enhanced Document-Query Interactions. *ArXiv180901682 Cs*. Published online September 11, 2018. Accessed November 16, 2020. http://arxiv.org/abs/1809.01682
318. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016. http://www.deeplearningbook.org
319. Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*. Association for Computational Linguistics; 2008:1070. doi:10.3115/1613715.1613855
320. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. *J Biomed Inform*. 2018;77:34-49. doi:10.1016/j.jbi.2017.11.011
321. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560
322. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229-236. doi:10.1136/jamia.2009.002733
323. Friedman C. Towards a comprehensive medical language processing system: methods and issues. *Proc Conf Am Med Inform Assoc AMIA Fall Symp*. Published online 1997:595-599.
324. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying Respiratory Findings in Emergency Department Reports for Biosurveillance using MetaMap. Published online 2004:5.
325. Liu M, Shah A, Jiang M, et al. A study of transportability of an existing smoking status detection module across institutions. *Am J Physiol - Ren Fluid Electrolyte Physiol*. 2012;2012:577-586.
326. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc JAMIA*. 2009;16(3):328-337. doi:10.1197/jamia.M3028
327. Soysal E, Wang J, Jiang M, et al. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc*. 2018;25(3):331-336. doi:10.1093/jamia/ocx132
328. Zheng K, Vydiswaran VGV, Liu Y, et al. Ease of adoption of clinical natural language processing software: An evaluation of five systems. *J Biomed Inform*. 2015;58:S189-S196. doi:10.1016/j.jbi.2015.07.008
329. Bompelli A, Silverman G, Finzel R, et al. Comparing NLP Systems to Extract Entities of Eligibility Criteria in Dietary Supplements Clinical Trials Using NLP-ADAPT. In: Michalowski M, Moskovitch R, eds. *Artificial Intelligence in Medicine*. Lecture Notes in Computer Science. Springer International Publishing; 2020:67-77. doi:10.1007/978-3-030-59137-3_7
330. Montani S, Striani M. Artificial Intelligence in Clinical Decision Support: a Focused Literature Survey. *Yearb Med Inform*. 2019;28(1):120-127. doi:10.1055/s-0039-1677911

331. Ozaydin B, Berner ES, Cimino JJ. Appropriate use of machine learning in healthcare. *Intell-Based Med*. 2021;5:100041. doi:10.1016/j.ibmed.2021.100041
332. Bohle S. "Plutchik": artificial intelligence chatbot for searching NCBI databases. *J Med Libr Assoc JMLA*. 2018;106(4):501-503. doi:10.5195/jmla.2018.500
333. Shneiderman B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Bederson BB, Shneiderman B, eds. *The Craft of Information Visualization*. Interactive Technologies. Morgan Kaufmann; 2003:364-371. doi:10.1016/B978-155860915-0/50046-9
334. Yang T, Chen Y, Emer J, Sze V. A method to estimate the energy consumption of deep neural networks. In: *2017 51st Asilomar Conference on Signals, Systems, and Computers*. ; 2017:1916-1920. doi:10.1109/ACSSC.2017.8335698
335. Rath A, Salamon V, Peixoto S, et al. A systematic literature review of evidence-based clinical practice for rare diseases: what are the perceived and real barriers for improving the evidence and how can they be overcome? *Trials*. 2017;18(1):556. doi:10.1186/s13063-017-2287-7
336. Chen Y, Mani S, Xu H. Applying Active Learning to Assertion Classification of Concepts in Clinical Text. *J Biomed Inform*. 2012;45(2):265-272. doi:10.1016/j.jbi.2011.11.003
337. Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw G. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *Int J Inf Manag*. 2018;38(1):1-6. doi:10.1016/j.ijinfomgt.2017.08.002
338. Bedford-Petersen C, Weston SJ. Mapping Individual Differences on the Internet: Case Study of the Type 1 Diabetes Community. *JMIR Diabetes*. 2021;6(4):e30756. doi:10.2196/30756
339. Gabarron E, Årsand E, Wynn R. Social Media Use in Interventions for Diabetes: Rapid Evidence-Based Review. *J Med Internet Res*. 2018;20(8):e10303. doi:10.2196/10303
340. Forge D. Algorithmic Bias in Machine Learning. In: Gordon and Betty Moore Foundation; 2019.
341. Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y. Measuring Bias in Contextualized Word Representations. *ArXiv190607337 Cs*. Published online June 17, 2019. Accessed October 14, 2021. http://arxiv.org/abs/1906.07337
342. Bordia S, Bowman SR. Identifying and Reducing Gender Bias in Word-Level Language Models. *ArXiv190403035 Cs*. Published online April 5, 2019. Accessed October 14, 2021. http://arxiv.org/abs/1904.03035
343. Baeza-Yates R. Data and algorithmic bias in the web. In: *Proceedings of the 8th ACM Conference on Web Science*. WebSci '16. Association for Computing Machinery; 2016:1. doi:10.1145/2908131.2908135
344. Statista, DataReportal. Active social network penetration in selected countries and territories as of January 2021. Published 2021. Accessed February 11, 2021. https://www.statista.com/statistics/282846/regular-social-networking-usage-penetration-worldwide-by-country/
345. Alammar J. The Illustrated Transformer. Published 2018. Accessed July 30, 2021. http://jalammar.github.io/illustrated-transformer/

ANNEXES

ANNEX 1. Active social network users in selected countries 2021

(Figure on next page)

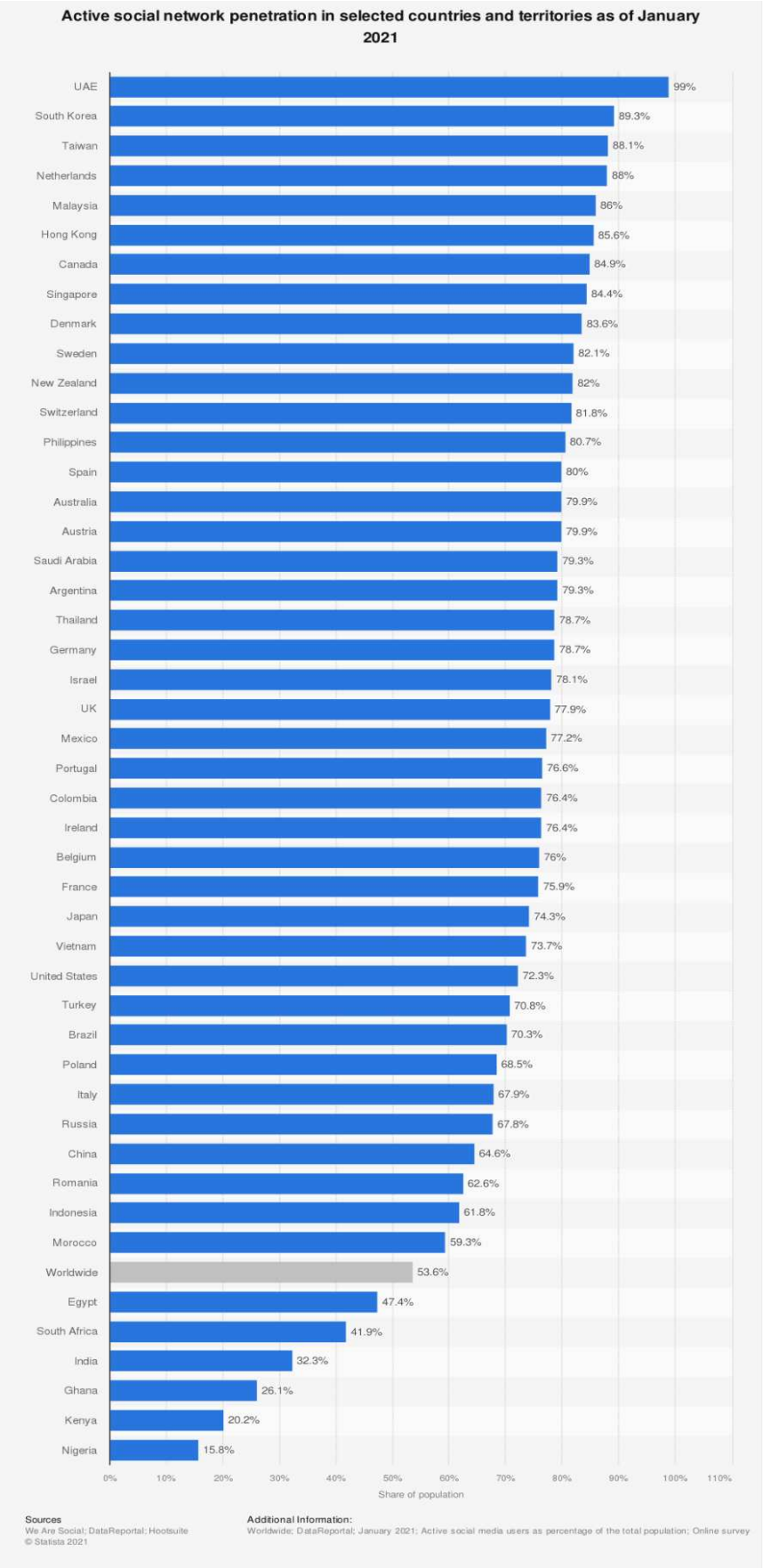


Figure A.1 Active social network users in selected countries 2021 (Figure below).[344]

ANNEX 2. Attention mechanism and transformer architecture

The attention mechanism was proposed by Bahdanau[187] and is a way of selectively weighting different words in the input such that they will have an impact on the hidden states resulting in a weighted context vector that weights the outputs of all previous prediction steps. When the model processes a given word, the self-attention mechanism looks at other positions in the input sentence which might help to better encode the word. Alammar provides a great self-attention illustration,[345] taking the example:

"The animal didn't cross the street because **it** was too tired"

Self-attention allows us to answer the question, to what "it" refers to, compare Figure A.2. When the model processes the word "it", self-attention associates "it" with "animal" emphasizing that the impact of "animal" is more important for "it" than other words like "because" or "was".

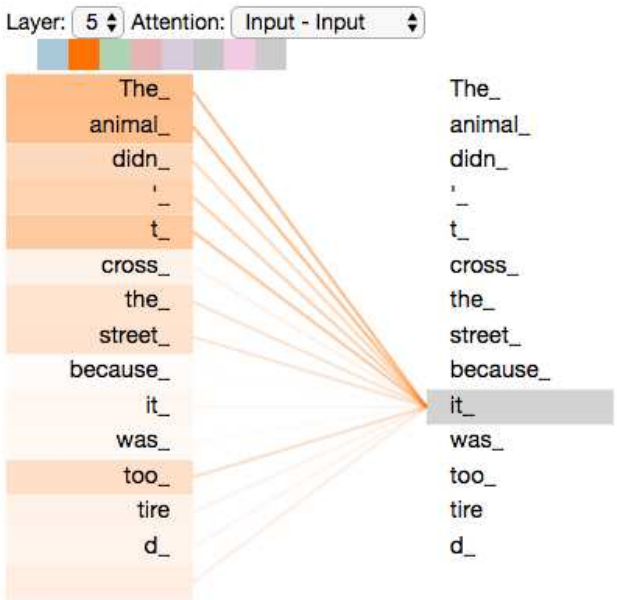


Figure A.2: Self-attention visualisation shows that during the encoding in the 5th encoder, the word "it" received strong attention from "The" and "animal".[345]

The authors described an attention function as "mapping a query and a set of key-value pairs to an output, where the query, keys, values, and outputs are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key".[65] The position encoded input vectors are split

into three flows: the queries (Q), keys (K) and values (V), which are obtained by multiplying the input embeddings by three matrices which are trained during the training process. The attention output in the original Transformer is “Scaled Dot-Product Attention” calculated by:

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

where the queries and keys have dimension d_k and values dimension d_v.

To further improve performance, *Multi-headed* attention was introduced which creates multiple sets of Query, Key, Value weight matrices, each of them randomly initialized, which each projects the input embedding into a different representation subspace. By learning to look at a text sequence “from different angles”, the Transformer becomes more context-aware. The authors used as default $h = 8$ attention heads. The multiple attention heads are then concatenated and multiplied by another weight matrix W^O , which is trained during the training process to produce the output matrix:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O,$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V),$$
$$W_i^Q \in \Re^{d_{model} \times d_k}, W_i^K \in \Re^{d_{model} \times d_k}, W_i^V \in \Re^{d_{model} \times d_v}, W_i^O \in \Re^{hd_v \times d_{model}},$$
$$d_k = d_v = 64,\ d_{model} = 512$$

The transformer architecture exploits the attention mechanism while processing sequences in parallel. All words are handled simultaneously rather than word-by-word.
More technically, the entire transformer architecture is a encoder-decoder mechanism leveraging scaled dot-product attention and multi-head attention as illustrated in Figure A.3.

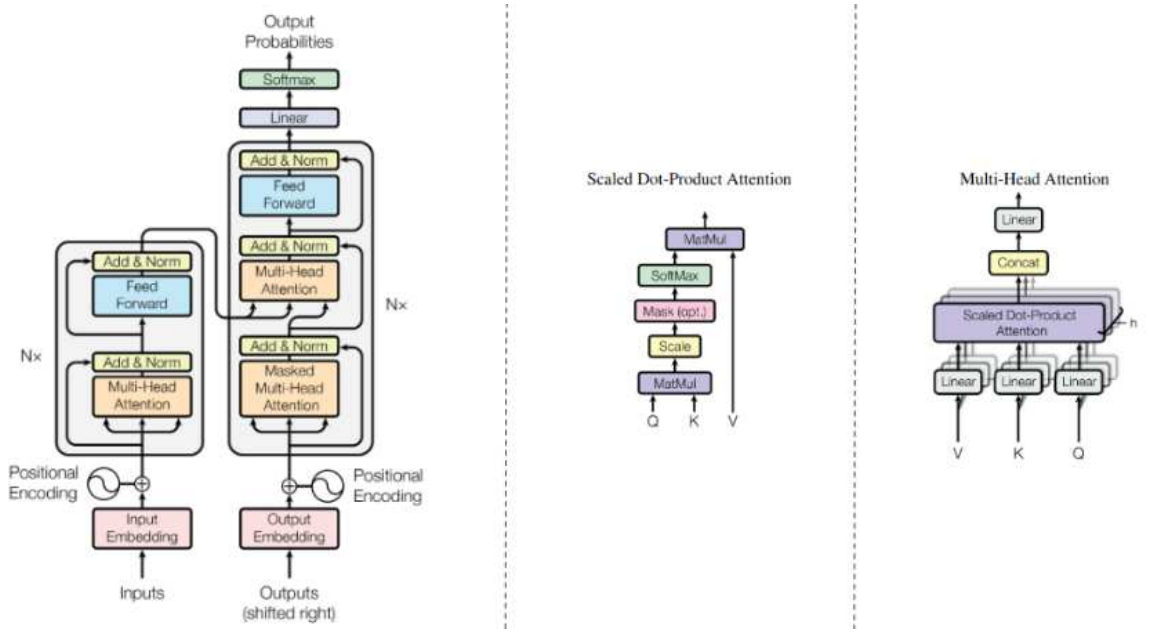


Figure A.3: Transformer architecture with encoder and decoder (left); Scaled Dot-Product Attention (center); Multi-Head Attention consists of several attention layers running in parallel (right).[65]

Figure A.3 on the left provides a global overview over the architecture with the encoder part on the left side and the decoder part on the right side. The input embedding converts tokens into a vector format and combines them with positional encodings, adding positional information. This combined vector enters the encoder which is composed of the following components: 1) multi-head attention mechanism which performs self-attention over a sequence to determine for each token which other tokens of the sentence are relevant/attend to that token; 2) fully connected feed-forward network which encodes the token after the multi-head attention in a d_{model}- dimensional vector; 3) residual connection allowing the model to optimize more efficiently because gradients can flow freely from the end to the beginning of the model; 4) add & norm block merges the output of the multi-attention head or feed forward layer with the residual. In the original proposition of the Transformer the authors stacked 6 encoder layers on top of each other. Those steps create an intermediate representation of the input which will then be translated into output predictions by the decoder which consists of: 1) the tokenized output is vectorised and combined with the positional encoding to obtain the output embedding for the decoder; 2) masked multi-head attention which applies multi-head self-attention on the output embedding while masking future positions to only attend to earlier positions in the output sequence; 3) another multi-head attention block using the combination of the encoded inputs from the encoder with the outputs of the prior masked multi-head attention to enable the model to learn correlations of the encoded inputs and

desired outputs; 4) feed forward; 5) A final linear fully connected network projecting the decoder output vector to logits vector of the length of the vocabulary; 6) A last softmax layer turns the logits into probabilities resulting in a probability for each word. Taking the argmax on this vector provides the predicted word.

In the training process, the true output sequence and the predicted output sequence are fed into a loss function to obtain an *error* which is used to backpropagate the neural network and update the parameter weights.

ANNEX 3. Primary, secondary and tertiary emotions, as defined by Parrot

Primary emotion	Secondary emotion	Tertiary emotion
Joy	Cheerfulness	Amusement · Bliss · Gaiety · Glee · Jolliness · Joviality · Joy · Delight · Enjoyment · Gladness · Happiness · Jubilation · Elation · Satisfaction · Ecstasy · Euphoria
	Zest	Enthusiasm · Zeal · Excitement · Thrill · Exhilaration
	Contentment	Pleasure
	Pride	Triumph
	Optimism	Eagerness · Hope
	Enthrallment	Enthrallment · Rapture
	Relief	Relief
Love	Affection	Adoration · Fondness · Liking · Attraction · Caring · Tenderness · Compassion · Sentimentality
	Lust/Sexual desire	Desire · Passion · Infatuation
	Longing	Longing
Surprise	Surprise	Amazement · Astonishment
Sadness	Suffering	Agony · Anguish · Hurt
	Sadness	Depression · Despair · Gloom · Glumness · Unhappiness · Grief · Sorrow · Woe · Misery · Melancholy
	Disappointment	Dismay · Displeasure
	Shame	Guilt · Regret · Remorse
	Neglect	Alienation · Defeatism · Dejection · Embarrassment · Homesickness · Humiliation · Insecurity · Insult · Isolation · Loneliness · Rejection
	Sympathy	Pity · Mono no aware · Sympathy
Anger	Irritability	Aggravation · Agitation · Annoyance · Grouchy · Grumpy · Crosspatch
	Exasperation	Frustration
	Rage	Anger · Outrage · Fury · Wrath · Hostility · Ferocity · Bitterness · Hatred · Scorn · Spite · Vengefulness · Dislike · Resentment
	Disgust	Revulsion · Contempt · Loathing
	Envy	Jealousy
	Torment	Torment
Fear	Horror	Alarm · Shock · Fear · Fright · Horror · Terror · Panic · Hysteria · Mortification
	Nervousness	Anxiety · Suspense · Uneasiness · Apprehension (fear) · Worry · Distress · Dread

Figure A.4: Primary, secondary and tertiary emotions. The colors in the table correspond to the distribution of primary emotions in Table 4.9.[79] [233]

ANNEX 4. Sample tweets for each topic

No.	Topic label	Sample tweets (closest to cluster center by cosine distance)
1	Support/solidarity in diabetes community	- Happy birthday to my T1D brother 🐕 ❤️ 🎁 🎉 🎃 🎂 USER hope your day is a great one 👊 HTTPURL - HAPPY BIRTHDAY TO MY FAVORITE DIABETIC 💔 I'm glad we're becoming close friends, come back soon I miss you :(I love you … - Closing: It was great getting to know everyone tonight! Thanks USER and Happy belated Birthday! #dsma
2	Inspiring relatives living with diabetes	- My amazing godson is 1 of these T1D warriors.He's tough as can be& filled w awesomeness!Sorry I couldn't walk today - USER What a talented family! You were amazing yesterday. Thank you for being a great example for living beyond T1D diagnosis. - Love the profile pic. Thanks for being such a great advocate and an inspiration to every person living with T1D!
3	Sharing hope and encouraging	- USER 2019. Hope, always hope, for better things. Plan to do the work to get there. #dsma - USER as a fellow t1d, your tweet made me smile. hope you get to say you no longer need batteries one day soon :) - USER hope all gets well, would like to donate but got my own diabetes bills as well. <3 hope she stays strong and lives long
4	Diabetes awareness / Support / Donation	- **ATTENTION** Please come out; support T1D that USER is hosting! Help us find a cure for this disease 💙 - Was told today that my work for people w/diabetes is just a passion project & only b/c I have T1D. Ask me for a favor ever again. Please. - What I like: When most people see something bad happen, they try to help. No insulin for kiddo? Here's a donation!!
5	DSMA* enjoying online support	- Good evening everyone! Welcome to #dsma! How are feeling tonight? #dsma - Closing: As always lots of love to everyone in the #DOC. Have a great week and remember that a little heart can do big things. #dsma - Q5 #DSMA maybe ask how #s make us feel and what can we do about it? I've done a lot of tweaking. All alone. Doc knows I'm good at it. #DSMA
6	Sharing diabetes related stories	- Q1: I have a really hard time being content. I'm in a good place right now with most things in life but still antsy for new adventures. #dsma - good morning i'm awake ONLY to give USER his insulin however i am also wishing USER a good day thank you for coming to my ted talk. - USER Oh good I'm glad it went well. Sorry you're still in pain. I have unfortunate news, I am for sure Diabetic type 2 Hypoglasimic
7	T1D hashtags	- A little bit of awesome (a lot a bit of awesome). #t1Dlookslikeme #tandem #diabetes #dexcom #g6 #t1D #jdrf #labrat #T1D - Awesome morning! Woke with a 🦄 for #muffinswithmom ! #pirrungelementary . #t1d #type1life #Type1Conqueror - Yesterday was a good day!! #t1d #dexcom #t1dpastor #caffeinatedpastor #gooddaywithdiabetes
8	Diabetes care	- My doctor says we must treat Depression like Diabetes, take care every day. Even when you feel good, you gotta stay…... URL - Taking care of your mental health is important. But dont eat for the first time in 3 days then make a thread about ur struggle w diabetes :/ - USER Your pancreas probably pumped out a lot of insulin, to take care of that load; 'good thing this is a rare practice for you 😉
9		- First rant ever about #bloodsugar palette. Palette is amazing but situation

	Bloodsugar palette (beauty products)	around it SUCKS terribly. #JeffreeStarCosmetics - Finally was able to get my #bloodsugar by @JeffreeStar! Now I just need my #blueblood palette - @JeffreeStar is coming out with the new #thirsty palette and I'm still trying to get my hands on the #BloodSugar palette! 😫😫😩
10	Advocacy for affordable insulin	- #insulin4all Amazing to see how far this movement has come that USER threads about insulin pricing. Started HTTPURL - I am THRILLED to see all of the publicity the insulin crisis is getting— let's keep it coming #insulin4all - Who said healthcare activism can't also be a hella good party ?! #DanceParade #Insulin4all see you in Albany
11	Life with and without diabetes	- I dream what my life would be like without Type 1 Diabetes, I wouldn't be depressed. I wouldn't sick all the time or worry. - Strange to think that not too long ago, kids that developed Type 1 diabetes would just get sick and die and the parents wouldn't know why... - sometimes I wish I had a best friend or someone I knew closely with diabetes, because they would understand what I go through
12	Life with type 1 diabetes	- 2nd day ever wearing my #Dexcom and it fails. It should last 10 days... You could say I'm a little mad 😠 #t1d #diabetes #dexcomg6 - Self esteem is at an all time low today. Send virtual love. Really need it right now! #diabetes #type1 #fuckme - Wish my blood sugars would sort itself out so i can enjoy my time with my gorgeous daughter.. #worrying #t1d #diabetic #newborn #babyaria 💔
13	Glucose guardian	- USER USER I need a glucose guardian 😢 - USER is my glucose guardian and I love her. - sometimes i wish i had a glucose guardian 😔
14	Chatting about insulin	- Love seeing my family can't afford insulin and Cory Booker could out-freestyle Bernie Sanders on the timeline. Someone please kill me.. - Made a manic thought come true now I have a sweet looking insulin pin used heart tattoo. Sorry mom.. 😞 - It was rly good and my insulin tubing was safe one bird tricked me and got my finger but it wasn't too bad, it was a lot of fun!
15	Insulin and insulin pump complications	- It doesn't take five minutes, but I really hate changing my insulin pump site and refilling the insulin reservoir. - Today: no fast-acting insulin left. Got sick, it lasted all day long. Used all spoons going to get insulin. Could not get long-acting insulin. - I think I stuck the insulin catheter in a bad spot this morning. Insulin isn't really having an effect. ugly
16	Diabetes in family	- With diabetes an issue for many, this is a real concern. Most people know at least 1 friend, family member, or coworker. - As someone who comes from a family who are high risk of getting diabetes, this is ridiculous and sad :/. - @fuxwidri My mom was with one of my sisters and diabetes runs on her side of the family 😕 so I'm worried
17	Diaversary	- got diagnosed with type 1 diabetes 16 years ago today, let's get fucked upp. Happy st pattys 🤪 - Happy Diabirthday Kid. 11 years since dx. Remember when 'they' said there would be a cure in 5 years? tick...tock… HTTPURL - Good morning world, today is a very personal day for me 21 years ago I was diagnosed with Type 1 Diabetes
18	Day-to-day stories about diabetes	- Today is gonna be a good day cus I get to see @ellieraen AND I'm getting a new insulin pump!! It's been a shitty week I need this!!! - YES! I remembered to test and take my morning insulin!! 130 is a good

		reading fir me. All my diabetic family out - Been taking Ben on my walks. Tonight I felt listless, weak so we came home. Take my glucose reading...75. Too low. Glad he was with/me in case
19	Confusion between T1D and T2D	- Dorit. Type 2 diabetes doesn't "turn" into type 1. And a 318 glucose reading is absolutely horrible-& - #diabetes I swear. I have type 1, and nothing is more frustrating than when people try to give me some type 2 lecture - Last night I watched What The Health and got really annoyed about their lack of distinction between Type 1/Type 2 diabetes. #NotAllDiabetes
20	Diagnosis	- I was just told that my dad's friend's 4 year old son was diagnosed with Type 1 Diabetes; they want me to go talk to him & my heart broke 😟 - The last time he showed me affection was 2007. I got diagnosed with Type 1 Diabetes and he said That sucks, sorry - my 12 year old son was diagnosed with type 1 diabetes on monday and i am terrified for him and his health but also what about his insurance..
21	Glycemic instability	- Every time my blood sugar is good I think I can eat whatever without giving insulin then an hour later I'm 350 with moderate ketones.. 😅 - I started consuming glucose gel whatever I feel weak at my work and it really boosts me up and I'm wondering if I have low blood sugar - Testing my blood glucose all day. Happy to say that right now, it is at 91. I'm gonna do my damnedest to keep my glucose normal. #diabetes
22	Insulin	- Makes me sad to discard insulin at work when i know there are type I diabetics who use expired insulin because they can't afford treatment😭 - how much insulin they have in their body at any given time. It also makes it less awkward to take insulin for a meal. - People say Type II diabetes is insulin resistance. I think it means your body doesn't produce enough insulin. Feel like I'm doomed if wrong.
23	Diabetes pop star Nick Jonas	- i remember being so little and finding out nick jonas was diabetic. i cried, i thought he was gonna die - I still remember that I cried at a Jonas brothers concert while nick was giving his speech about living with diabetes - Were you really a jonas brothers fan if you never cried over Nick Jonas having diabetes at some point in elementary school
24	Misunderstandings of diabetes	- USER we R getting the run around on our daughters T1D monitor; pump. Along w/ Dexacom. Makes real hard 2 believe u care. Am I wrong? - Idk if living alone with uncontrolled type one diabetes was a good idea. Hell even if it was controled you just never know. - USER Very good question.. Because the next.step is saying let's not treat t2 diabetes.. Because people ate too much.
25	Frustration with insulin prices	- My insulin is $1800. Utterly ridiculous! I can't afford mine either and sometimes I have to go to an urgent care or ER for insulin. - I know the fear of people who take insulin. I am one of them. Even with insurance, insulin is so expensive. - I think this is gonna be the first year I don't get VIP USER upgrades. When my insulin costs $500 a pop I can't justify the cost. sad :/
26	OGTT**	- USER Yeah i had my regular glucose test today, and they are having me come back tomorrow to do the 3 hour glucose test :/ - I haven't felt the baby move since i did my glucose test and i hate to be one of those "i think the glucose test kills - Have to take my glucose test today and it's been the only thing I've been dreading this whole pregnancy.

27	Parents and diabetes	- My mom is like so embarrassed sometimes of my sister having to take her insulin in public. Woman it's normal just get over it - Today, I was diagnosed with diabetes. My mom texted “I’m sorry, babygirl”. I cried because I’ve seen her fight with this my whole entire life - So sad 2 hear a friend of family with type 1 diabetes died. She was only 19yo. She was so sweet & full of life. Glad she knew Jesus as savior
28	Diabetes Distress	- T1D hasn't been my 1st concern lately, but the days where I'm reminded oh, this can definitely kill me, I can't help feel a little' depressed - I know it's all I talk about but I hate this. I hate being diabetic so much I feel so contricted. Ranting about it makes it hurt less sorry :(- Wow I really hate being diabetic. I’m like 99% sure my last pen of insulin was bad cause I have been feeling like shit
29	Diabetic/Insulin Shock	- I went into shock after a cortisone shot (potentially insulin shock because I'm hypoglycemic/ pre-diabetic) - Diabetic shock at 23. I went into shock after eating three Klondike bars - USER Someone went into diabetic shock or something when we were by like the North Pole. Crazy. She was ok by the end of the flight ..
30	Diabetes-related comorbidities	- My birth mom is diabetic (even missing toes now), has gastroparesis, liver/heart issues, chronic pain, and more. Her meds alone are immense. - “Diabetic nerve damage take this it’ll stop that. but could make you have severe depression, suicidal thoughts and...” I’ll keep the pain - people who feel the need to constantly one up you. like oh? you have depression? well i have anxiety, type 2 diabetes, and arthritis

Table A.1: Sample tweets for each topic. ** OGTT : Oral glucose tolerance test
User mentions were replaced by the token “USER” and urls by “HTTPURL”. Tweets were slightly adapted to ensure privacy.

ANNEX 5. Top emotional keywords and emojis/emoticons by topics of interest of people living with diabetes

No.	Topic label	top emotional words	top emojis / emoticon
1	Support / solidarity in diabetes community	happy, love, hope, good, awesome, glad, celebration, sunshine	💙, ❤️, :), 😂, 😍, 🤗, 🙃, 💜
2	Inspiring relatives living with diabetes	love, amazing, awesome, excited, good, happy, glad, hope	💙, ❤️, :), 😂, 😊, 😍, 💕, 💜
3	Sharing hope and encouraging	hope, good, feel, love, promise, happy, glad, understand	❤️, :(, 😔, 😊, 😂, 💖, 😟, 💗
4	Diabetes awareness / Support / Donation	love, good, hope, happy, feel, loved, attention, understand	💙, ❤️, :), 😂, 😃, 🤗, 🙃, 🙄
5	DSMA enjoying online support	good, glad, love, hope, feel, awesome, happy, excited	:), ❤️, 😂, dx, :(, 💙, ;), 😊
6	DSMA sharing diabetes-related stories	good, bad, feel, love, glad, happy, awesome, hope	:), 😂, ❤️, 🤗, 🤗, 😊, 🙃, 😢
7	T1D hashtags	love, happy, good, awesome, excited, hope, amazing, hate	💙, ❤️, 😂, :), 😍, 😕, 💜, 🤣
8	Diabetes care	care, good, love, feel, hope, bad, glad, sick	😂, ❤️, 🙄, 🤗, 🙃, 😔, 💜, 🤒
9	Bloodsugar palette (beauty products)	excited, love, happy, amazing, sad, loving, bad, feel	❤️, 😍, 🤗, 💙, ♥, 💕, 💋, 🤒
10	Advocacy for affordable insulin	good, love, hope, amazing, awesome, glad, attention, rally	💙, 💜, 😢, :), 🤗, :3, 😱, ❤️
11	Life with and without diabetes	love, good, feel, hope, bad, understand, surprised, happy	😂, ❤️, 🤗, dx, 🙄, :), 🙃, 😢
12	Life with type 1 diabetes	love, good, feel, bad, happy, hate, excited, hope	💙, ❤️, 😂, 🤗, 🙄, :), 🙃, 🤒
13	Glucose Guardian	love, good, happy, care, bad, poor, hope, feel	😂, 🤗, 🤒, ❤️, 🤣, 😔, 🙄, 🙃
14	Chatting about insulin	love, good, feel, hope, bad, happy, funny, hate	😂, ❤️, :), 🤗, :(, 🙃, dx, 🙄
15	Insulin pump complications	love, hope, good, feel, bad, excited, happy, hate	😂, 🙃, 🤗, 🙄, :), 🤒, ❤️, 😑
16	Diabetes in the family	good, love, hope, bad, feel, care, suffer, understand	❤️, 😂, 💙, 😊, 🙄, 😔, 🤗, 😢
17	Diaversary	good, happy, feel, love, bad, care, sick, hope	😂, ❤️, dx, :), 🙄, 🙃, 😊, 💙
18	Day-to-day stories about diabetes	good, feel, bad, happy, sick, love, hate, excited	😂, 🤗, 🙄, 🙃, ❤️, 🤒, :3, :)
19	Confusion between T1D and T2D	good, love, feel, bad, understand, hope, sick, hate	💙, 🤗, ❤️, 😂, 🙄, 🙃, :(, :)
20	Diagnosis	sick, love, good, feel, happy, care, bad, hope	dx, ❤️, 🤗, 💙, 😂, 😢, :(, 😡
21	Insulin	good, feel, bad, understand, love, hope, hate, sick	😂, 🙄, 🤗, 🙃, ❤️, :), :(, 😑
22	Glucose variability	good, bad, feel, love, hate, sick, happy, understand	😂, 🤗, 🙃, ❤️, 🙄, 🤒, :), ♥
23	Diabetes pop star Nick Jonas	cried, crying, love, cry, crush, feel, good, sad	🤗, 😂, :(, 💜, 😑, :/, 🤣, 😱
24	Misunderstandings of diabetes	understand, love, bad, good, feel, hate, sick, sad	😂, 🙄, 🤗, ❤️, 😡, 🙃, 🤒, 😢
25	Frustration with insulin prices	good, care, outrageous, sad, love, hope, bad, feel	🙄, 😡, 🤗, 😂, 🙃, 😢, :(, 😑
26	OGTT	bad, good, hate, feel, nervous, dreading, excited, hope	🤗, 🤒, 🙄, 🤢, 🙃, 😂, 😑, 😬
27	Parents and diabetes	bad, good, sick, love, sad, poor, care, feel	😂, 🤗, ❤️, 😲, 😑, :(, 🙄, 😔
28	Diabetes Distress	feel, hate, sick, bad, sad, worried, funny, glad	😂, 🤗, :(, 🙄, 😢, :), ❤️, 😔
29	Diabetic / Insulin shock	shock, love, worried, good, bad, horrible, sweetness, trauma	🙄, 😂, 🤒, 🙃, 😑, 😲, 😊, 😱
30	Diabetes-related comorbidities	pain, depression, anxiety, bad, feel, suffer, good, damage	❤️, 🙃, 🤗, 😂, 🙄, :(, 😲, 😔

Figure A.5: Top emotional keywords and emojis/emoticons by topics of interest. [79]

ANNEX 6: Annotation guidelines

Objective

The aim of this labeled corpus is to provide a training data set for detecting possible cause-effect-pairs in diabetes and diabetes distress related tweets. Diabetes distress regroups all psychological factors related to the day-to-day disease management such as emotional burden, stress, anxiety, emotions, etc.

Data

Between April 2017 and January 2021 diabetes related tweets have been extracted using the Twitter API based on list of diabetes-related keywords, such as “diabetes”, “insulin”, “hypoglycemia”, “#T1D”, “#DSMA”, “Type 2”, “#diabeteslookslikeme”, compare Supplementary File 1 for the full list. Based on the extracted tweets a random subsample of 5000 tweets has been selected for annotation purposes.

File structure - Columns:

- **Text [String]**: extended tweet message “Full_text“ in tweet object
- **Intent [String]**: Intent of the tweet. If several intents, they are separated by a semicolon (“;”) Can take the following values:
    - q: Question in tweet
    - mS: multiple sentences in tweet
    - mC: multiple causes in tweet
    - mE: multiple effects in tweet
    - msS: multiple sentences in tweet with cause and effect in a single sentence
    - neg: A negation which negating the meaning of the cause or effect in a sentence
    - joke: A joke, an irony or sarcasm is in the tweet

- **Cause [String]:** Words describing the causes. A cause can be composed of several words. If several causes occur in a tweet then they are separated by a semicolon (“;”)

- **Effect [String]:** Words describing the effect. An effect can be composed of several words. If several effects occur in a tweet then they are separated by a semicolon (";")
- **Causal association [0,1]**: Binary variable of a cause-effect pair occurs in a tweet where 0 means no cause-effect pair and 1 means there is a cause-effect pair

Definition of a cause-effect relationship and annotation rules

The following tweet examples are fictive to ensure privacy.

Non-diabetes or diabetes distress related relationships

The focus on this corpus lies on cause-effect relationships related to diabetes and diabetes distress. For this reason, tweets like the following are not labeled as causal. The possible cause here might be "flu" and the effect "die", but "flu" is out of scope in this project.

Tweet	Intent	Cause	Effect	C.A.*
Scary, i have a 13 year old diabetic daughter however i read 4 thousand or more people a year die in UK just from flu. so why this fuss & panic over corona . I read lots and had nightmares last night ! ! This is ridiculous	mS			0
Schools are closed to prevent passing the virus , yet ALL DAY LONG they are in the store with parents , putting me and MY HEALTH at risk ! WHY, WHY, WHY?				0

* causal association

In the second example "heart disease" is not labeled as the *cause* as it is out of scope.

Examples for possible causal associations

Tweet	Intent	Cause	Effect	C.A.*
Diabetes causes me to have mood swings. :/		Diabetes	mood swings	1
years of diabetes and all I got is a Spidey-sense like ability to notice any abnormal sensations in my body and about 7 new kinds of anxiety I didn't know existed before I got diagnosed . it sucks , but I'm much stronger because of it .	msS;mE	diabetes	abnormal sensations in my body;anxiety	1
Gestational Diabetes is shit . The poking , urge to eat the foods that are no good for me . When I am in need of more insulin my body alarms are all going off , making me tired , headache , blurred vision .	msS;mE	in need of more insulin	tired;headache; blurred vision	1
After 10 years of injections and finger pricks, I have finally gotten an Insulin pump and glucose monitor . Finally I can start to manage my diabetes even better and improve my health . Waited so long	mS;mC;mE	Insulin pump;glucose monitor	manage my diabetes even better;improve my health	1

* causal association

The above examples also show that several causes can lead to an effect, and inversely also one cause can lead to several effects.

Implicit relations

The cause-effect relationship is not stated by a *causal link* word

Tweet	Intent	Cause	Effect	C.A.*
I was sent to the penalty box to fix a low blood sugar #diabetes #NHLPlayoffs		#diabetes	low blood sugar	1

* causal association

Unclear cause - effect relationships

In tweets in which there is a possible cause and a possible effect but it is not clear if the “cause” had an influence on the “effect”, the tweet is labeled as non-causal.

Tweet	Intent	Cause	Effect	C.A.*
I'm back ! Had two strokes recovering now my legs do not want to move. Have high blood pressure and diabetes . So all of you out there please watch you're blood pressure . It matters .				0
My dad has diabetes, cancer , heart problems , and a weak immune system .				0

* causal association

The possible cause is “High blood sugar and diabetes” and the possible effect is “stroke”. But it can not be concluded that the stroke was provoked by the high blood sugar or diabetes

Several chaining cause-effect relationships: A -> B -> C

If in a tweet we have two relationships: event A causes event B and at the same time event B causes event C, then we labelled the relationship that is closest to our objective to study diabetes and diabetes distress:

Tweet	Intent	Cause	Effect	C.A.*
Not sure if I've been up since 3:30 for Titan or because my anxiety over my glucose test is keeping me up 😬 Bahh		glucose test	anxiety	1
I am also a diabetic with all this worry & stress , is adding to my sugar levels to rise ..	mE	diabetic	worry;stress	1
Excess insulin from eating too many carbs spikes insulin , making you hungry. Belief me		eating too many carbs	Excess insulin;spikes insulin	1

* causal association

event A: glucose test
event B: anxiety
event C: been up since 3:30

Diabetes Distress

In labeling this dataset, a special focus was lying on diabetes distress (psychological factors related to the day-to-day disease management, such as anxiety, stress, emotions, etc.). For this reason we labeled possible causal associations related to diabetes distress as well:

Tweet	Intent	Cause	Effect	C.A.*
I do I just want to go to the kitchen and eat . I hate #diabetes	msS	#diabetes	hate	1
I have gestational diabetes and im very much bothered		gestational diabetes	bothered	1
Kent ' s just angry because his diabetes is flaring up again .		diabetes	angry	1

* causal association

Negations

If a negation word occurs in a cause or effect, it is considered being part of the cause and effect and so not altering the meaning of the cause or effect. Consequently the tweet is not labeled as having a negation “neg” in the Intent.

Tweet	Intent	Cause	Effect	C.A.*
My 14 year old daughter . Type 1 (malfunctioning pancreas , aka not enough insulin being made to regulate	msS;mE	Type 1	malfunctioning pancreas; not enough insulin	1
I'm a Type 1 Diabetic , out of work and unable to afford my insulin 😓		out of work	unable to afford my insulin	1
I was wondering why i felt like shit and then I realized I haven't given myself my insulin since early this morning . stupid..	neg	insulin	felt like shit	1
Don't hate your diabetes ; instead , find ways to love it and get rid of it over time . it helps	neg	diabetes	hate	1
my friend " gave " herself diabetes by not doing what her doctor told her LOSE WEIGHT !	neg	LOSE WEIGHT	diabetes	1

* causal association

The last example shows when a tweet is labeled as negation. The negation “haven’t given myself” alters the meaning of the causal relationship “insulin” -> “felt like shit”.

Jokes

As jokes were also labeled tweets containing ironic or sarcastic elements.

Tweet	Intent	Cause	Effect	C.A.*
This tweet is so dumb it gave me diabetes	joke			0
I love lifestyle choices become a non smoker and a temporary diabetic , why the hell not *irony out*	joke			0
Thanks sweetie And I think I've developed diabetes from your sweetness	joke			0

* causal association

Frequently used abbreviations related to diabetes

Abbreviation	Explanation
lows, going low	low blood sugar
#gbdoc, #doc, #dsma	diabetes related online groups on social media to exchange about the disease
dexcom, Freestyle Libre	continuous glucose monitoring tools helping to monitor blood sugar levels levels
cgm, CGM	continuous glucose monitoring
DKA, dka	Diabetic Ketoacidosis
3 hours	3 hour glucose test for gestational diabetes
LCHF	low carb high fat diet: The diet, because of its low requirement for insulin, has been recognised by the Swedish government as being suitable for people with type 2 diabetes and as helpful to individuals looking to lose weight or maintain a healthy weight.
BS	blood sugar

ANNEX 7: Most frequent cause/effect clusters

The synonyms were manually added for the initial clusters. Clusters whose parent cluster is "Other", are automatically added clusters that were predicted from the cause-effect classifier.

Most frequent clusters				
N°	**Parent cluster**	**cluster**	**synonyms**	**N**
0	Diabetes	diabetes	diabetic, #diabetic, #diabetes, diabetes mellitus, diabetics, DIABETIC, #Diabetic	66775
1	Death	death	die, passed away, gave life, kill, killing, dead, shorter lifespan, loose live, died, dying, commit suicide, losing father	16989
2	Insulin	insulin	insulin hormone, supplies, HUMALOG bottle	14148
3	Diabetes	type 1 diabetes	T1D, type 1, #type1, #type1diabetes, juvenile diabetes	11693
4	Emotions	fear	anxiety, terrified, scared, anxious, concern, dread, scary, worried, nervous, distress, panic, hysteria, terrified, creeped, traumatized	10160
5	Glycemic variability	hypoglycemia	low glucose, low blood sugar, low, hypo, go low, glucose down, blood sugar down, lowered BS, sugar drop	9547
6	Symptoms	sick	headache, dizzy, threw up, vomiting, painful, puke, feeling sick, shaky, nausea, coughing, diarrhea, dry mouth, fever, sweat	6549
7	Nutrition	overweight	obese, eat too much, weight gain, fat, gained pounds, obesity	5186
8	Diabetes	type 2 diabetes	T2D, type 2, #type2, #type2diabetes, T2 diabetic, t2d, TYPE TWO	4909
9	Complications & comorbidities	neuropathy	amputation, feet amputation, lost feet, lost leg, leg amputation, nerve death, nerve damage, lost hand, diabetes neuropathy	4481
10	Healthcare system	medication	meds, diabetes meds, drug, antibiotics, pills, drugs, dose, medicine, metformin prescribed, prescription, cheapest medicines	4389
11	Diabetes Technology	insulin pump	injecting insulin, injection, inject, needle, pump, finger prick, shot,	4307
12	Nutrition	nutrition	vegan, vegetarian, eat carbs, carbohydrates, no chocolate, can't eat donuts, food, salad, noodles, appetite fish, NEEDED meal, seafood, milk, entire meal, broccoli	4230
13	Emotions	anger	rage, outrageous, frustration, hate, angry, jealous, jealousy, raging, pissed, pissed off, frustrating	4149
14	Health	OGTT	glucose test drink, glucose test, 3 hour test, ogtt, diabetic drink, horrific drink	4053
15	Blood pressure	hypertension	high blood pressure, BP	3782
16	Healthcare system	finance	wages, student loans, GoFundMe, expenses, costly, expensive care, pay, price gouging, spend money, debt, insulin price, insulin prices, donations, Healthcare Cost	3767
17	Nutrition	reduce weight	lost pounds, lose weight, #loseweight	3589
18	Insulin	unable to afford insulin	can't afford insulin, no access to affordable insulin, could not afford insulin, can't afford meds, could not buy insulin, bankrupt, financially unstable	3381
19	Nutrition	diet	diabetic diet, Keto diet, carnivore diet, keto, plant based diet, high fat	3325

			diet, change diet, #lowcarb, LCHF, dietary needs, Low Carb	
20	Emotions	sadness	cry, sad, sucks, loneliness, lonely, sadly, CRIED, despair, hurtful, hurting, psychological grief, disappointing	3153
21	Glycemic variability	hyperglycaemia	high blood sugar, high glucose, high glucose levels, spike glucose, higher blood glucose, blood glucose up, blood glucose levels up, elevated #BP, rebellious #hyperglycemia	3144
22	Diabetes	suffer	suffering, TERRIBLE PAIN, HURT	3132
23	Diabetes Distress	depression	depressed, depressing, lose hope, mentally ill, hopeless, antidepressants, psychologically fragile	2810
24	Healthcare system	hospital	surgery, syringes, doctor, appointment, checkup, medical attention, medical treatment, ER, ICU, hospitalize, ambulance, doc, surgeries, GP practice, clinical psychologist, Caregiver	2721
25	Diabetes Distress	stress	mood disorder, stressed, stressful,, MOOD SWINGS, tense	2681
26	Nutrition	sugar	sweets, candy, waffle, soda, Sugar, CAKE, artifical sweeteners, CRAVE SWEETS, milkshakes, LOVE SUGAR	2369
27	Nutrition	fasting	starvation, not eating	2363
28	Insulin	rationing insulin	shortage insulin, denying insulin, lack insulin, ration insulin, EXPIRED INSULIN	2244
29	Health	gestational diabetes	pregnancy, pregnant	2076
30	Health	prediabetes	pre diabetic, borderline diabetic	1932
31	Diabetes Distress	feel bad	feel awkwared, disgusting, appetite, grumpy	1861
32	Complications & comorbidities	retinopathy	horrible vision, bad eyesight, vision decline, lost sight, blind, diabetes retinopathy	1750
33	Complications & comorbidities	high risk	risk	1663
34	Healthcare system	insurance	company, pharma, health insurance, coverage, Medicare, #BigPharma #Insulin, medicaid	1627
35	Complications & comorbidities	coma	unconscious, pass out, Diabetic Coma	1540
36	Complications & comorbidities	heart attack	cardiovascular, cardiovascular disease, diabetic heart attack, CHF	1511
37	Health	insulin resistance		1505
38	Complications & comorbidities	complications	diabetes complication, issues	1443
39	Other	struggle		1357
40	Complications & comorbidities	nephropathy	kidney damage, diabetes kidney failure, Nephrologist #Diabetes #Nephrologist,kidney failure	1338
41	Emotions	joy	feel good, feel better, relief, happy, proud	1279
42	Other	constant pain		1240
43	Other	n't know language		1234
44	Diabetes Distress	fatigue	no power, without energy, exhaused, tired, lethargic, burnout, exhaustion	1190
45	Pandemic	covid	corona, coronavirus, virus, covid pandemic business, Corona, vaccine, worrying COVID, severity COVID, VIRUS, #CoronavirusPandemic, SARS CoV 2 INFECTION, vaccinated	1073
46	Lifestyle	physical activity	exercising, walking, sport, exercises, walk, gym, gyms	1066

47	Healthcare system	politics	health system, NHS, #brexit, brexit, Canada, capitalism, government, EU, administration, frustrated #NHS, economy, CANADA, CAPITALISM	985
48	Insulin	access insulin	no insulin, don't have insulin, without insulin, #insulin4all, #Insulin4all #Diaversary	956
49	Diabetes Technology	continuous glucose monitor	freestyle libre, #freestylelibre, monitoring, #dexcom, CGM, cgm	934
50	Insulin	affordable insulin	afford insulin	915
51	Complications & comorbidities	shock	dIaBeTiC sHOcK	908
52	Pandemic	home	staying home, quarantine, shutdown	864
53	Diabetes	management	control diabetes, uncontrol	825
54	Complications & comorbidities	infection	wound, wounds, inflammation	822
55	Symptoms	Insomnia	can't sleep, awake, wake, sleepy, asleep	737
56	Health	lost job	without work, laid off	723
57	Complications & comorbidities	diabetic ketoacidosis	keto acidosis, #ketoacidosis, diabetic ketoacidosis, DKA, #ketoacidosis, keto acidosis, DAIBETIC KETO ACIDS high, #KETO #NSNG	710
58	Health	immune system		705
59	Nutrition	eating healthy	#healthy #meal	608
60	Blood pressure	hypotension	low blood pressure	587
61	Family	family	brother, daughter, grandpa, dad, mom, grandma, parent	581
62	Complications & comorbidities	renal failure	diabetes renal failure, dialysis	554
63	Complications & comorbidities	legs swollen	foot swelled	544
64	Diabetes	reverse diabetes	reversed, cured overnight	489
65	Diabetes community	support	#dsma, help, raise awareness, supporters	483
66	Lifestyle	lifestyle	environment	453
67	Emotions	love	like	451
68	Other	cold		443
69	Complications & comorbidities	cholesterol		392
70	Other	meat fake meat		376
71	Other	starve		363
72	Diabetes Distress	isolation	alone, live alone, distrust	358
73	Other	cut back rice		348
74	Other	taking necessary precautions		346
75	Complications & comorbidities	pancreas	diabetes pancreas	340
76	Health	PCOS	pcos, PCOs, Pcos	333
77	Other	ass alive		318
78	Other	arthritic		312
79	Other	dangerous		301

80	Other	vulnerable		299
81	Other	seizures		285
82	Other	acting		274
83	Other	needles sensation		273
84	Other	slightly concerned		267
85	Insulin	insulin spike	insulin jump	263
86	Complications & comorbidities	liver failure	diabetes liver failure	256
87	Other	fucked drunk		253
88	Complications & comorbidities	stomach		246
89	Symptoms	thirsty	thirst, dehydrated, THIRSTY	245
90	Other	stop talking		244
91	Nutrition	alcohol	beer, BEER, alcoholism	243
92	Glycemic variability	A1C	a1c, predict HbA 1c	236
93	Health	genetic	genes, Genetics, hereditary, genetically modified, shit genetics	228
94	Complications & comorbidities	cancer	chemo, #Cancer	226
95	Other	hair fall		226
96	Other	crash hard		212
97	Other	shut kidneys		211
98	Other	isolated ' society		210
99	Other	afraid ingredients		204
100	Other	surgeon excited		202

Table A.2: Most frequent cause/effect clusters

ANNEX 8: Hierarchical clustering formulas

To provide a more formalized description of this clustering process, a deep dive into the formulas is outlined in the following.

A clustering node is defined by the *head words* p_j^l of its children, where $j \in \{0, \ldots, N_h\}$ is the index of the *head word* and $l \in \{1,2\}$ specifies if the *head word* belongs to clustering child node 1 or 2. N_h is the number of *head words* for a given node, from which $N_h / 2$ are associated to each child. By default the number of *head words is* $N_h = 6$, resulting in *head words* p_0^1, p_2^1, p_4^1 for child 1 and p_1^2, p_3^2, p_5^2 for child 2. For simplicity reasons, we also refer to a *head word* with only the lower, unique index p_j or simply j.

A document traverses the tree and at each node, to decide to which child node the document goes, following steps are performed:

1) *Token scores:* Each token is compared to the *head words* of the two child nodes
2) *Children score:* The *token scores* are aggregated to calculate a score for each child which determines the future path of the document.
3) *Head word improvement*: Test if replacing an existing *head word* with a new token leads to improved representation of the documents having passed this node.

Token scores

Each token t_i of a given document performs at each node:

- Children creation: Create two new clustering children if the following conditions are satisfied:
    - children do not exist already
    - maximum number of nodes is not reached
    - the number of documents having arrived at this node is superior to the parameter *childSplitSize* (default: 50)
- Children initialisation: Test if each child has been initialised with $N_h / 2$ *head words.* Otherwise take first distinct tokens of the document to affect them simultaneously to both children until N_h is reached.
- Token scores: return the highest cosine similarity to a *head word* of child 1 and child 2:

$$score_{ij}^1 = max_j \, sim(t_i, p_j^1)$$

$$score_{ij}^2 = max_j\ sim(t_i, p_j^2)$$

After this step, each token is attributed to the *head word* he is most similar to in child 1 ($score_{ij}^1$) and the *head word* he is most similar to in child 2 ($score_{ij}^2$).

Children scores

The token scores are then aggregated to obtain a global score for each child: *children score*. The document will then follow its path to the child node with the highest *children score*:

- Aggregation all *token scores* per *head word* to obtain *head word scores*:

$$hwScore_j^1 = avg_{i\in\theta_j^1}\ score_{ij}^1 \qquad \text{and} \qquad hwScore_j^2 = avg_{i\in\theta_j^2}\ score_{ij}^2$$

$$\theta_j^l = \{i\text{: tokens } t_i \text{ who scored highest for a head word } p_j^l\}$$

- Aggregation of *head word scores* for each child:

$$childScore_1 = avg_{j\in\omega_1}\ hwScore_j^1$$
$$childScore_2 = avg_{j\in\omega_2}\ hwScore_j^2$$
$$\omega_l = \{j : \text{head words in child node } l\}$$

The document will then go to the child node with maximal *childScore*. Let's denote $\hat{l}$ the index of this child.

The last step which is performed for each document at each node is the *head word* improvement, that requires the definition of further internal variables, specific to each node, that are updated at the same time during each iteration:

- Weighted sum of *token score* to for the *head word* j:

$$pScores_j = pScores_j + \sum_{i\in\Omega_l} score_{ij}^l * w_l$$

where, $\Omega_l = \{i\text{: } score_{ij}^l > score_{ij}^{l'}\}$, are all tokens which are most similar to any *head word* in child l and l' being the other child;

and $w_l = \frac{1}{|\Omega_l|}$ is a weighting factor by the number of highest *token scores* for a *head word* in child l.

- Center word embedding for each *head word* j as weighted sum of all *token scores* for j:

$$vCenters_j = vCenters_j \frac{pScores_j}{pScores_j + w_j} + \sum_{i\in\Omega_j} t_i \frac{w_l}{pScores_j + w_l}$$

where $\Omega_j = \{i: score_{ij}^{l} > score_{ij'}^{l}\}$ are all tokens which are most similar to *head word* j, with j' being the other *head words* in the child l

- Center word embedding of all tokens that passed the current node:

$$center = \frac{\sum_j vCenters_j * pScores_j}{\sum_j pScores_j}$$

- Distance of the *head word* j embedding to the center of all tokens with highest *token score* for j:

$$pGAP_j = 1 - sim(vCenters_j, p_j)$$

The hypothesis here is, the smaller $pGAP_j$, the better p_j represents the tokens which scored highest for j

- Weighted sum of all $pGAP_j$ over all *head words* j:

$$GAP_k = \sum_j pGAP_j * (\frac{pScores_j}{\sum_j pScores_j})$$

This is a global (error) measure for the node k of how well the *head words* represent the traversed tokens.

Head word improvement

The last step consists in testing if one of the tokens with highest *token score* for a *head word* $p_j^{\hat{l}}$, from the child $\hat{l}$ the document got associated to, signifies an improved representation of the documents having passed this node. Specifically, a token t_i that achieved the highest *token score* for a *head word* $p_j^{\hat{l}}$, will replace respective *head word* $p_j^{\hat{l}}$ if two conditions are fulfilled:

- if t_i is closer to the center vector embedding of all tokens that scored highest for $p_j^{\hat{l}}$ than the *head word* itself:

$$sim(vCenters_j, t_i) > sim(vCenters_j, p_j^{\hat{l}})$$

This can be interpreted as if the token t_i would better represent those tokens have passed this node.

- if t_i is closer to the sum of all *head words* of the child to which the document goes than to the center of all tokens that passed through the node:

$$sim(center, t_i) < sim(\sum_{j \in \Omega_j} p_j, t_i)$$

This condition avoids that the potential new *head word* t_i is too general in terms of semantics and thus not discriminative enough as a *head word*.

Figure A.6 visualises those formulas with a simple example at the beginning of tree building. At the start of the tree building, the tree only consists of the "In Scope"-classifier and his clustering child node. The sentence "*Diseases are diabetes and cancer*" is the first to be fed to the tree. The "In Scope" classifiers classifies each token of being relevant based on the initial training dataset a user has specified. In this example the tokens "are" and "and" were classified as not relevant, as containing too little discriminative information. Only the tokens "Diseases", "diabetes" and "cancer" pass the node. To decide then to which of the two child nodes the sentence goes, each of the three tokens gets compared to the *head words* of both children.

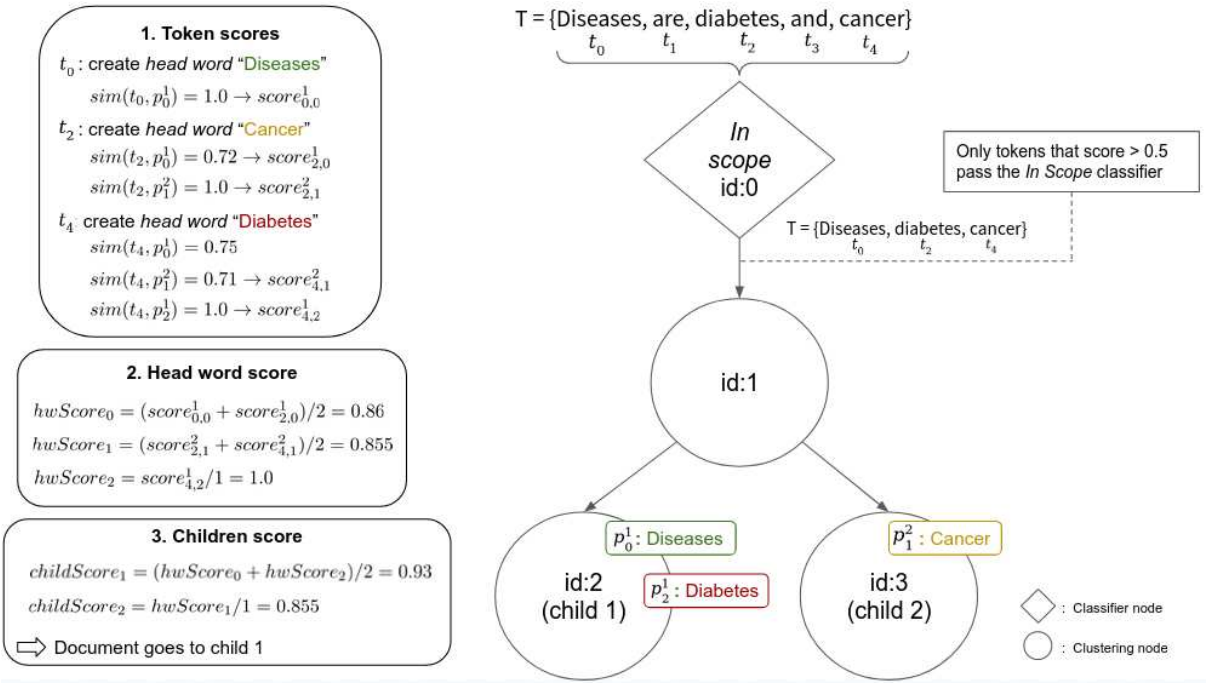


Figure A.6: Score calculations for the first steps at tree building without having been initialized with *head words*. Maximum number of *head words* per node of $N_h = 6$.

As it is the first iteration, the children get created and have no *head words* yet. The first token "Diseases" is then assigned to be the first *head word* p_0^1 in child 1 and only the *token score* of "Disease" to this single *head word* can be calculated, the $score_{0,0}^1$. The second token "diabetes" initialises the first *head word* p_1^2 in child 2 and can then be compared to the two *head words* leading to $score_{2,0}^1$ for child 1 and $score_{2,1}^2$ for child 2. The third token "cancer" gets then affected again to

child 1 as *head word* p_2^1. For a longer sentence the first distinct N_h tokens define the initial *head words*. The token "cancer" can then calculate the similarity to the three *head words* and only the highest cosine similarity to a *head word* in each child node is returned, in this case $score_{4,2}^1$ for child 1 and $score_{4,1}^2$ for child 2.

The next step consists in aggregating the *token scores* per *head word*, here 3 *head word scores*, which will then be aggregated to *childScores* in the last step. In the example the *childScore* for child 1 is higher causing the document to be assigned to child 1. As in this example the *head words* are not fully initialized, no *head word* improvement is applied.

Figure A.7 shows another example with already initialized *head words* in the child nodes. The "In Scope"-classifier predicts the tokens "pancreas" and "essential" to be the relevant tokens. Each token is then compared to the *head words* of both child nodes again to calculate the *token scores*.

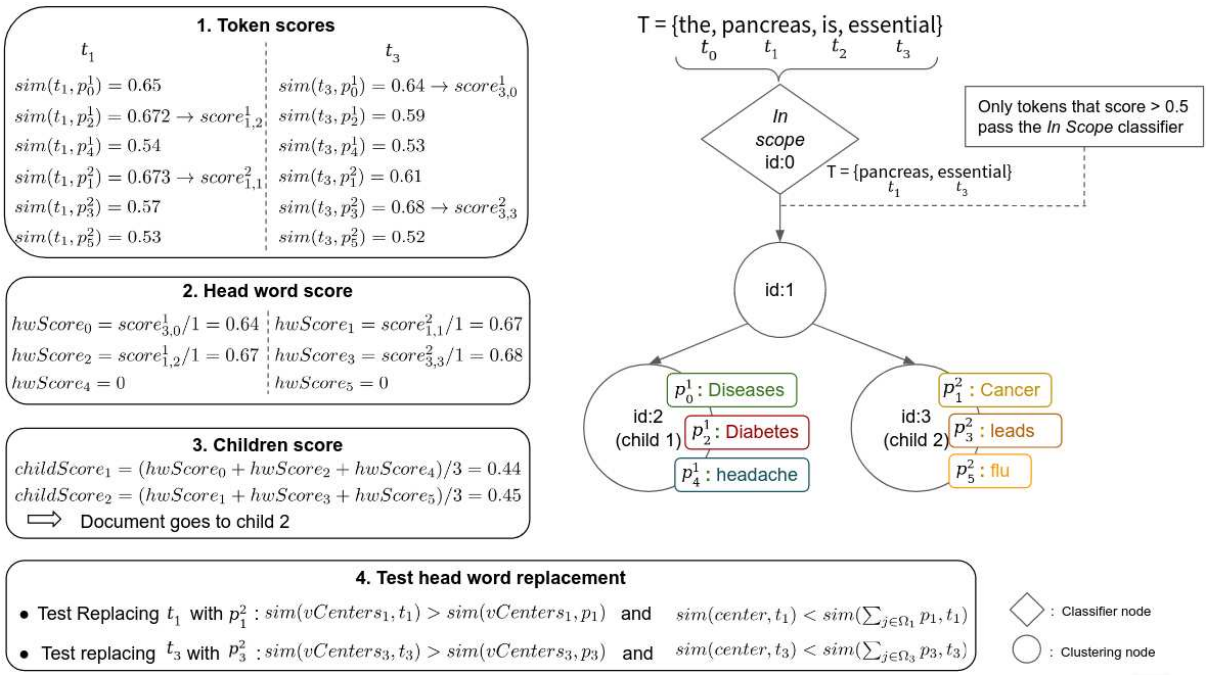


Figure A.7: Score calculations for an already with *head words* initialized child nodes. Maximum number of *head words* per node of $N_h = 6$.

Head word score aggregation, followed by the *childScore* calculation leads to the document being assigned to child 2. Token t_1 achieved highest *token score* for *head word* 1 and token t_3 scored highest for *head word* 3, both *head words* in child node 2. In consequence, both tokens are tested if replacing them by their closest *head words* brings an improvement in terms of *head word* improvement.

Once all documents have been processed in the tree, a global measurement *clusteringGAP*, the sum of all aggregated distances of *head words* to their closest tokens, over all nodes, is calculated:

$$clusteringGAP = \sum_{k} GAP_k$$

And a new iteration starts with documents being re-processed through the tree. The algorithm has stops when the error distance measure $clusteringGAP$ does not minimize further and a local minimum has been reached.

Intuitively, this convergence criteria states that the most representative tree structure is obtained when no more *head words* can be replaced in any node and thus the best possible *head words* have been found resulting in the highest possible level of interpretability.

ANNEX 9. Active learning performance for all MeSH codes

Pos. class	N°	Random						Uncertainty Sampling						FeedbackExplorer						CNN - Zhang					
		pos	neg	Acc	Prec	Rec	F1	pos	neg	Acc	Prec	Rec	F1	pos	neg	Acc	Prec	Rec	F1	pos	neg	Acc	Prec	Rec	F1
complications D048909	50	31	19	0.79	0.77	0.94	0.84	29	21	0.71	0.70	0.90	0.79	29	21	0.78	0.88	0.72	0.80	26	24	0.62	0.72	0.78	0.59
	100	67	33	0.73	0.69	0.98	0.81	39	61	0.78	0.77	0.91	0.83	61	39	0.83	0.88	0.82	0.85	57	43	0.73	0.81	0.74	0.76
	150	94	56	0.81	0.79	0.93	0.85	57	93	0.83	0.84	0.87	0.86	88	62	0.84	0.84	0.90	0.87	76	74	0.78	0.83	0.83	0.81
	200	125	75	0.81	0.79	0.93	0.85	78	122	0.85	0.91	0.82	0.86	110	90	0.85	0.85	0.91	**0.88**	104	96	0.81	0.84	0.85	0.84
Test set: 591 (pos) / 409 (neg) || Train set: 557 (pos) / 443 (neg)																									
Angiopathies D003925	50	14	36	0.86	0.78	0.48	0.60	11	39	0.84	0.86	0.30	0.45	17	33	0.88	0.80	0.53	0.64	10	40	0.79	0.02	0.33	0.03
	100	23	77	0.86	0.93	0.37	0.53	55	45	0.90	0.76	0.73	0.75	45	55	0.90	0.81	0.56	0.72	43	57	0.85	0.35	0.87	0.52
	150	35	115	0.85	0.93	0.33	0.48	76	74	0.88	0.67	0.86	0.75	65	85	0.89	0.75	0.75	0.75	69	71	0.87	0.54	0.85	0.63
	200	44	156	0.86	0.94	0.35	0.51	87	113	0.87	0.84	0.46	0.60	76	124	0.90	0.87	0.61	0.72	95	105	0.90	0.68	0.83	**0.74**
Test set: 209 (pos) / 791 (neg) || Train set: 178 (pos) / 822 (neg)																									
Foot D017719	50	2	48	0.93	0	0	0	3	47	0.95	0.62	0.51	0.56	6	44	0.95	0.72	0.34	0.46	2	48	0.93	0	0	0
	100	5	95	0.93	0	0	0	19	81	0.96	0.81	0.52	0.64	37	63	0.95	0.73	0.45	0.56	14	86	0.94	0.07	0.27	0.11
	150	8	142	0.93	0	0	0	42	108	0.96	0.77	0.61	0.68	50	100	0.95	0.74	0.45	0.58	40	110	0.94	0.39	0.64	0.42
	200	12	188	0.93	0	0	0	49	154	0.96	0.82	0.46	0.59	53	147	0.96	0.81	0.43	0.56	56	144	0.97	0.65	0.88	**0.73**
Test set: 67 (pos) / 933 (neg) || Train set: 59 (pos) / 941 (neg)																									
Retinopathy D003930	50	2	48	0.95	1.0	0.28	0.43	4	46	0.98	0.92	0.79	0.85	11	39	0.96	0.68	0.88	0.77	3	47	0.92	0	0	0
	100	8	92	0.97	0.94	0.58	0.72	30	70	0.98	0.84	0.92	0.88	20	80	0.98	0.92	0.86	0.88	21	79	0.95	0.37	0.87	0.46
	150	10	140	0.96	0.93	0.51	0.66	33	117	0.97	0.93	0.67	0.78	44	106	0.98	0.89	0.86	0.87	54	96	0.98	0.83	0.92	0.87
	200	11	189	0.96	0.97	0.46	0.63	51	149	0.98	0.93	0.74	0.82	63	137	0.98	0.92	0.78	0.84	62	138	0.98	0.83	0.83	**0.87**
Test set: 76 (pos) / 924 (neg) || Train set: 68 (pos) / 932 (neg)																									
Cardiomyopathies D058065	50	1	49	0.95	0.24	0.27	0.25	3	47	0.97	0.5	0.03	0.06	1	49	0.97	0	0	0	2	48	0.97	0	0	0
	100	3	97	0.97	0	0	0	15	85	0.97	0.56	0.47	0.51	3	97	0.97	0	0	0	3	97	0.97	0	0	0
	150	5	145	0.97	0	0	0	23	127	0.97	0	0	0	5	145	0.97	0	0	0	7	143	0.95	0.02	0.10	0.03
	200	7	193	0.97	0	0	0	26	174	0.97	0	0	0	13	187	0.97	0	0	0	12	188	0.97	0.01	0.20	**0.02**
Test set: 30 (pos) / 970 (neg) || Train set: 28 (pos) / 972 (neg)																									
Coma D003926	50	1	49	0.95	0	0	0	1	49	0.95	0	0	0	5	45	0.95	0.67	0.04	0.08	2	48	0.95	0	0	0
	100	4	96	0.95	0	0	0	28	72	0.93	0.37	0.62	0.46	13	87	0.95	1.0	0.04	0.08	13	87	0.95	0.05	0.20	0.07
	150	5	145	0.95	0	0	0	38	112	0.96	0.62	0.32	0.42	25	125	0.95	0.8	0.08	0.15	33	117	0.97	0.43	0.89	0.56
	200	7	193	0.95	0	0	0	45	155	0.96	0.86	0.12	0.21	29	171	0.95	1.0	0.04	0.08	46	154	0.97	0.44	0.99	**0.59**
Test set: 50 (pos) / 950 (neg) || Train set: 50 (pos) /950 (neg)																									
HHNK* D006944	50	1	49	0.98	0	0	0	3	47	0.96	0.11	0.08	0.10	1	49	0.98	0	0	0	1	49	0.98	0	0	0
	100	4	96	0.98	0	0	0	7	93	0.98	0.33	0.04	0.07	4	96	0.98	0	0	0	5	95	0.98	0	0	0
	150	6	144	0.98	0	0	0	16	134	0.98	0	0	0	6	144	0.98	0	0	0	14	136	0.98	0.01	0.20	0.02
	200	6	196	0.98	0	0	0	22	178	0.98	0	0	0	7	193	0.98	0	0	0	21	179	0.98	0.05	0.48	**0.08**
Test set: 24 (pos) / 976 (neg) || Train set: 26 (pos) / 974 (neg) * Hyperglycemic Hyperosmolar Nonketotic Coma																									
Ketoacidosis D006944	50	1	49	0.98	0	0	0	3	47	0.96	0.11	0.08	0.10	1	49	0.98	0	0	0	2	48	0.96	0	0	0
	100	4	96	0.98	0	0	0	7	93	0.98	0.33	0.04	0.07	4	96	0.98	0	0	0	5	95	0.96	0	0	0
	150	6	144	0.98	0	0	0	16	134	0.98	0	0	0	6	144	0.98	0	0	0	11	139	0.96	0	0	0
	200	6	194	0.98	0	0	0	22	178	0.98	0	0	0	7	193	0.98	0	0	0	23	177	0.96	0	0	0
Test set: 38 (pos) / 962 (neg) || Train set: 31 (pos) / 969 (neg)																									

Nephropathies D003928	50	3	47	0.94	0.88	0.10	0.18	6	44	0.93	0.64	0.99	0.17	2	48	0.92	0.25	0.04	0.07	4	46	0.93	0.0	0.1	0.01
	100	9	91	0.94	0.79	0.15	0.26	39	61	0.83	0.26	0.80	0.39	11	89	0.94	0.93	0.18	0.31	15	85	0.93	0.08	0.4	0.13
	150	14	136	0.93	1.0	0.04	0.08	40	110	0.94	0.65	0.46	0.54	34	116	0.95	0.85	0.32	0.47	36	114	0.94	0.18	0.74	0.26
	200	16	184	0.93	0	0	0	52	148	0.95	0.83	0.35	0.50	55	145	0.95	0.76	0.49	**0.60**	55	145	0.95	0.33	0.97	0.48
Test set: 71 (pos) / 929 (neg) || Train set: 76 (pos) / 924 (neg)																									
Neuropathies D003929	50	7	43	0.90	0.76	0.45	0.57	9	41	0.69	0.30	0.92	0.45	8	42	0.89	0.80	0.32	0.45	6	44	0.86	0	0	0
	100	16	84	0.91	0.84	0.44	0.58	13	87	0.91	0.83	0.47	0.60	28	72	0.92	0.88	0.52	0.65	30	70	0.89	0.36	0.85	0.44
	150	23	127	0.91	0.88	0.44	0.59	47	103	0.94	0.81	0.72	0.76	60	90	0.93	0.81	0.69	0.74	60	90	0.90	0.63	0.78	0.64
	200	31	169	0.91	0.88	0.40	0.55	69	131	0.92	0.89	0.51	0.65	84	116	0.94	0.87	0.69	0.77	85	115	0.94	0.81	0.80	**0.80**
Test set: 139 (pos) / 861 (neg) || Train set: 137 (pos) / 863 (neg)																									
Fetal Macrosomia D005320	50	4	46	0.95	0.43	0.95	0.59	4	46	0.94	0.39	0.79	0.52	6	44	0.97	0.62	0.79	0.69	2	48	0.96	0	0	0
	100	8	92	0.97	0.64	0.76	0.70	12	88	0.98	0.88	0.67	0.76	18	82	0.98	0.69	0.81	0.75	11	89	0.97	0.24	0.56	0.32
	150	10	140	0.98	0.68	0.76	0.72	26	124	0.98	0.85	0.69	0.76	28	122	0.98	0.88	0.67	0.76	25	125	0.97	0.34	0.97	0.48
	200	12	188	0.97	0.84	0.38	0.52	29	171	0.98	0.97	0.5	0.65	30	170	0.98	0.96	0.57	**0.72**	31	169	0.97	0.26	0.99	0.38
Test set: 42 (pos) / 958 (neg) || Train set: 32 (pos) / 968 (neg)																									
Gestational D016640	50	2	48	0.93	0.48	0.24	0.32	5	45	0.93	0.49	0.63	0.55	3	47	0.94	0.58	0.54	0.56	3	47	0.93	0	0	0
	100	5	95	0.95	0.67	0.43	0.53	23	77	0.94	0.51	0.64	0.57	20	80	0.95	0.65	0.69	0.67	12	88	0.94	0.01	0.27	0.11
	150	8	142	0.94	0.71	0.07	0.14	31	119	0.95	0.70	0.45	0.55	36	114	0.95	0.61	0.78	0.68	37	113	0.94	0.39	0.64	0.42
	200	14	186	0.94	0.71	0.15	0.25	54	146	0.95	0.66	0.61	0.64	52	148	0.95	0.60	0.75	0.67	54	146	0.97	0.65	0.88	**0.73**
Test set: 67 (pos) / 933 (neg) || Train set: 62 (pos) / 938 (neg)																									
Experimental D003921	50	3	47	0.94	0.57	0.55	0.56	8	42	0.91	0.39	0.75	0.51	7	43	0.94	0.69	0.14	0.23	4	46	0.94	0	0	0
	100	8	92	0.94	0.51	0.72	0.59	18	82	0.94	0.75	0.18	0.30	22	78	0.94	0.89	0.12	0.22	30	70	0.94	0.24	0.70	0.28
	150	11	139	0.93	0.48	0.74	0.58	51	99	0.94	0.52	0.88	0.66	39	111	0.95	0.72	0.28	0.4	58	92	0.95	0.50	0.72	0.55
	200	15	185	0.94	0.66	0.29	0.40	61	139	0.95	0.65	0.57	**0.61**	58	142	0.95	0.76	0.34	0.47	73	127	0.96	0.53	0.78	0.59
Test set: 65 (pos) / 935 (neg) || Train set: 84 (pos) / 916 (neg)																									
Type 1 D003922	50	8	42	0.90	0.85	0.90	0.31	4	46	0.89	0.51	0.16	0.24	5	45	0.89	0.6	0.05	0.10	6	44	0.89	0.01	0.2	0.01
	100	14	86	0.89	1.0	0.02	0.03	23	77	0.92	0.72	0.49	0.58	16	84	0.89	0.9	0.08	0.14	23	77	0.89	0.14	0.71	0.19
	150	18	132	0.89	1.0	0.02	0.03	35	115	0.91	0.82	0.28	0.42	27	123	0.90	1.0	0.10	0.17	38	112	0.9	0.14	0.77	0.23
	200	22	178	0.89	0.80	0.03	0.07	58	142	0.91	0.80	0.29	**0.42**	44	156	0.90	0.78	0.22	0.34	53	147	0.91	0.19	0.98	0.30
Test set: 115 (pos) / 885 (neg) || Train set: 108 (pos) / 892 (neg)																									
Wolfram D014929	50	3	47	0.93	0.25	0.96	0.39	1	49	0.98	0.5	0.04	0.08	2	48	0.95	0.32	0.88	0.47	1	49	0.98	0	0	0
	100	4	96	0.98	0.67	0.67	0.67	17	83	0.96	0.34	0.88	0.49	6	94	0.96	0.34	0.83	0.49	9	91	0.98	0.12	0.4	0.16
	150	6	144	0.98	1.0	0.33	0.5	18	132	0.99	0.83	0.79	0.81	11	139	0.99	1.0	0.58	0.74	22	128	0.98	0.31	0.8	0.44
	200	6	194	0.98	1.0	0.04	0.08	24	176	0.99	0.94	0.63	0.75	20	180	0.99	0.89	0.71	**0.79**	27	173	0.99	0.43	1.0	0.59
Test set: 24 (pos) / 976 (neg) || Train set: 28 (pos) / 972 (neg)																									
Type 2 D003924	50	5	45	0.73	0.15	0.45	0.23	13	37	0.83	0.23	0.41	0.29	5	45	0.88	0.14	0.07	0.09	5	45	0.91	0	0	0
	100	11	89	0.9	0.27	0.08	0.12	28	72	0.91	0.33	0.02	0.04	18	82	0.91	0	0	0	14	86	0.91	0.01	0.05	0.02
	150	17	133	0.91	0.5	0.03	0.06	48	102	0.91	1.0	0.01	0.02	25	125	0.90	0.25	0.05	0.08	30	120	0.91	0.04	0.33	0.06
	200	22	178	0.91	0.0	0.0	0.0	73	127	0.90	0.41	0.28	**0.33**	31	169	0.91	0	0	0	51	149	0.92	0.08	0.53	0.14
Test set: 88 (pos) / 912 (neg) || Train set: 118 (pos) / 882 (neg)																									

Lipoatrophic D003923	50	4	46	0.98	0.33	0.33	0.33	2	48	0.98	0.33	0.06	0.10	1	49	0.98	0	0	0	1	49	0.98	0	0	0
	100	5	95	0.98	0	0	0	8	92	0.98	0	0	0	3	97	0.98	0	0	0	5	95	0.98	0.01	0.1	0.01
	150	6	144	0.98	0	0	0	26	124	0.99	1.0	0.17	0.29	18	132	0.99	1.0	0.17	0.29	16	134	0.98	0.03	0.18	0.06
	200	6	194	0.98	0	0	0	31	169	0.98	1.0	0.11	**0.20**	21	179	0.98	0	0	0	28	172	0.98	0.11	0.06	0.18
Test set: 18 (pos) / 982 (neg) || Train set: 33 (pos) / 967 (neg)																									
Donohue D056731	50	1	49	0.98	0	0	0	1	49	0.98	0.5	0.17	0.25	1	49	0.97	0.25	0.33	0.29	1	49	0.98	0	0	0
	100	3	97	0.98	0.36	0.22	0.28	12	88	0.98	0.44	0.67	0.53	9	91	0.99	1.0	0.33	0.5	7	93	0.98	0.1	0.30	0.14
	150	4	146	0.98	0	0	0	19	131	0.99	1.0	0.22	0.36	15	135	0.97	0.32	0.72	0.44	16	134	0.98	0.28	0.79	0.27
	200	5	195	0.98	0	0	0	21	179	0.98	1.0	0.06	0.11	18	182	0.98	1.0	0.06	0.11	19	181	0.99	0.27	0.8	**0.38**
Test set: 18 (pos) / 982 (neg) || Train set: 21 (pos) / 979 (neg)																									
LADA** D0000716 98	50	1	49	0.99	1.0	0.18	0.31	1	49	0.98	0	0	0	1	49	0.99	0.5	0.18	0.27	1	49	0.99	0	0	0
	100	1	99	0.99	0	0	0	3	97	0.99	1.0	0.18	0.31	3	97	0.99	1.0	0.18	0.31	1	99	0.99	0	0	0
	150	1	149	0.99	0	0	0	4	146	0.99	0	0	0	4	146	0.99	0	0	0	2	148	0.99	0	0	0
	200	2	198	0.99	0	0	0	4	196	0.99	0	0	0	5	195	0.99	0	0	0	3	197	0.9	0	0	0
Test set: 11 (pos) / 989 (neg) || Train set: 5 (pos) / 995 (neg) ** Latent Autoimmune Diabetes in Adults																									
Prediabetic D011236	50	2	48	0.96	0.5	0.02	0.04	1	49	0.96	0	0	0	1	49	0.96	0	0	0	2	48	0.96	0	0	0
	100	3	97	0.96	0	0	0	11	89	0.96	0.67	0.13	0.22	3	0.97	0.96	0	0	0	3	97	0.96	0	0	0
	150	7	143	0.96	0	0	0	27	123	0.96	0.75	0.07	0.12	9	141	0.96	0.67	0.04	0.08	9	141	0.96	0.	0.10	0.01
	200	8	192	0.96	0	0	0	31	169	0.96	0	0	0	16	184	0.96	0.53	0.18	**0.27**	17	183	0.96	0	0	0
Test set: 45 (pos) / 955 (neg) || Train set: 45 (pos) / 955 (neg)																									

Figure A.8: Performance of all MeSH codes for all four active learning strategies.[294] Information about the number of positive and negative training examples, the accuracy, precision, recall and F1-Score are provided.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Développement de méthodes d'intelligence artificielle pour l'analyse de données de réseaux sociaux et à des fins de recherche médicale : Cas d'utilisation sur une étude mondiale sur le diabète

Mots clés : diabète, épidémiologie numérique, médias sociaux, intelligence artificielle, aide à la décision clinique, causalité

Résumé : **Contexte** : Le diabète et la détresse liée au diabète représentent un fardeau mondial et leur incidence est en constante augmentation. L'épidémiologie traditionnelle du diabète présente plusieurs lacunes qui pourraient être comblées avec certaines approches innovantes. En effet, cela peut prendre de nombreuses années entre l'identification et la conception d'une question de recherche, l'obtention de la validation des autorités et l'inclusion des participants aux résultats de la recherche. L'épidémiologie numérique offre ainsi une opportunité de récolter rapidement des données en croissance exponentielle dans l'espace numérique. Il s'agit d'une source de données qui n'est pas disponible dans un contexte traditionnel. En outre, les systèmes d'aide à la décision clinique basés sur l'IA ont le potentiel d'aider les professionnels de la santé à filtrer les informations essentielles dans la masse de données textuelles disponibles telles que les dossiers de santé électroniques, la littérature scientifique ou les réseaux sociaux. Les **objectifs** principaux de cette thèse étaient 1) l'exploration des réseaux sociaux, comme source de données complémentaire pour l'épidémiologie du diabète; 2) le développement et l'open-sourcing de méthodes innovantes d'intelligence artificielle pour extraire des informations; 3) et fournir un système d'aide à la décision clinique aidant les professionnels de la santé à analyser les données textuelles en constante augmentation.

Résultats : Les principales préoccupations et sujets d'intérêt liés au diabète ont été identifiés, avec les émotions associées, mettant en lumière des sujets préoccupants sur l'accès aux soins, comme par exemple la frustration liée au prix de l'insuline aux États-Unis. Des associations "cause-effet" liées au diabète ont également été identifiées et visualisées dans un réseau interactif. Enfin, un système d'aide à la décision clinique interactif alimenté par une méthode d'intelligence artificielle a été développé pour améliorer l'exploration de la littérature dans le processus de prise de décision clinique, permettant une interprétabilité accrue tout en réduisant la consommation de mémoire. **Conclusion** : Ce travail a démontré que les données en ligne peuvent être utiles et complémentaires à celles de l'épidémiologie traditionnelle. Avec le cas d'usage du diabète, ce travail a également souligné l'importance des facteurs psychologiques et des émotions dans le quotidien et leur poids dans le fardeau de la maladie. Ce travail suggère une plus grande inclusion de ces dimensions dans les futures études épidémiologiques sur le diabète. Enfin, le besoin d'outils d'aide à la décision pour la pratique clinique pour synthétiser la littérature sur un sujet donné a été identifié et le prototype développé doit désormais être testé en situation réelle.

Title : Development of artificial intelligence methods for the analysis of online data for medical research purposes: Use case on the World Diabetes Distress Study

Keywords : diabetes, digital epidemiology, social media, artificial intelligence, clinical decision support, causality

Abstract : **Background:** Diabetes and diabetes distress represent a global burden and their incidence is constantly rising. Traditional diabetes epidemiology has several gaps that could be filled with certain innovative approaches. Indeed, it can take many years to identify and design a research question, acquire ethical approval, include participants and finally obtain research results. Digital epidemiology offers an opportunity to quickly harvest exponentially growing data in the digital space, a data source that is not available in traditional settings. In addition, AI-powered clinical decision support systems have the potential to assist health professionals filter critical information from the mass of available textual data such as electronic health records, scientific literature or social media. **Objectives:** The main objectives of this thesis were 1) the exploration of social media as complementary data source for diabetes epidemiology; 2) the development and open-sourcing of innovative artificial intelligence methods to extract information; 3) and to provide a clinical decision support systems helping health professionals to analyze the constantly growing clinical text data.

Results: Key diabetes related concerns and topics of interest were identified, along with associated emotions shared, highlighting areas of concern about access to care, such as the frustration concerning insulin prices in the US. Diabetes-related "cause-effect" associations have been identified and visualised in an interactive network. Lastly, an AI-powered interactive clinical decision support system has been developed to improve the literature exploration in the clinical decision making process enhancing interpretability while reducing memory consumption. **Conclusions:** This work demonstrated that online data can be useful and complementary to traditional epidemiology. Along with the example of diabetes, this work also highlighted the importance of psychological factors and emotions in everyday life and their weight in the burden of the disease. This work recommends a greater inclusion of these dimensions in future epidemiological studies on diabetes. Finally, the need for decision supporting tools for clinical practice to synthesize the literature on a given subject has been identified and the developed prototype must now be tested in a real scenario.