



**HAL**  
open science

# Argumentation quality : from general principles to healthcare applications

Santiago Marro

► **To cite this version:**

Santiago Marro. Argumentation quality : from general principles to healthcare applications. Artificial Intelligence [cs.AI]. Université Côte d'Azur, 2023. English. NNT : 2023COAZ4085 . tel-04402347

**HAL Id: tel-04402347**

**<https://theses.hal.science/tel-04402347v1>**

Submitted on 18 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# THÈSE DE DOCTORAT

Qualité de l'argumentation : Des principes  
généraux aux applications dans le domaine de  
la santé

**Santiago MARRO**

WIMMICS, Inria, CNRS, I3S

**Présentée en vue de l'obtention  
du grade de docteur en Informatique  
d'Université Côte d'Azur  
Dirigée par : Serena VILLATA**

**Soutenue le : 8 Novembre 2023**

**Devant le jury, composé de :**

**Président du Jury :** Maurizio FILIPPONE,  
Professeur des Universités, EURECOM

**Rapporteurs :**

Véronique MORICEAU, Maître de  
conférences HDR, Paul Sabatier Toulouse  
University 3

Frédérique SEGOND, Professeur des  
Universités, HDR, Inria



# **Qualité de l'argumentation : Des principes généraux aux applications dans le domaine de la santé**

**Argumentation Quality: From General Principles to Healthcare Applications**

## **COMPOSITION DU JURY**

**Président du jury:**

Maurizio FILIPPONE, Professeur des Universités, EURECOM

**Rapporteurs:**

Véronique MORICEAU, Maître de conférences HDR - Paul Sabatier  
Toulouse University 3

Frédérique SEGOND, Professeur des Universités, HDR, Inria

**Directeurs de thèse:**

Serena VILLATA, Directeur de Recherche, HDR, Université Côte d'Azur

**Invité:**

Elena CABRIO, Professeur des Universités, HDR, Université Côte d'Azur



# Résumé

L'analyse automatisée de l'argumentation a suscité un intérêt considérable ces dernières années, car les méthodes informatiques permettent d'améliorer la qualité du discours dans tous les domaines. Ceci est particulièrement pertinent dans des domaines complexes tels que les soins de santé, où un raisonnement solide a un impact direct sur la vie humaine. Le travail présenté dans cette thèse fait progresser l'état de l'art en matière d'extraction d'arguments et d'évaluation de la qualité, adapté aux complexités du domaine médical. La thèse apporte quatre contributions principales : (1) Développement et application des techniques d'extraction d'arguments, dont l'analyse de leur utilisation dans divers domaines et les contributions à la recherche COVID-19. (2) Méthodes d'évaluation de la qualité de l'argumentation, dont l'annotation d'un nouvel jeu de données de 402 essais d'étudiants avec des dimensions de qualité telles que la cohérence, la rhétorique et la vraisemblance. Des architectures neuronales innovantes combinant des caractéristiques textuelles et des encastresments de graphes se révèlent capables de classer correctement ces aspects, obtenant respectivement 0,78 F1, 0,89 F1 et 0,54 F1. (3) Identification des prémisses potentielles dans le domaine médical en analysant automatiquement les symptômes de 314 cas cliniques et en les alignant sur des sources de connaissances externes telles que l'ontologie du phénotype humain (HPO) à l'aide d'enchâssements contextuels (précision de 0,53). (4) Développement d'une fonction de prévalence transparente pour classer le pouvoir explicatif des prémisses identifiées, en s'appuyant sur des statistiques telles que l'anormalité et l'unicité de la base de connaissances. Cette thèse apporte des contributions significatives aux domaines de l'extraction d'arguments et de l'évaluation de la qualité grâce au développement de nouvelles techniques et ressources. Les méthodes proposées repoussent les limites de l'analyse automatique des arguments, tandis que les ensembles de données spécialement conçus offrent de nouvelles opportunités pour la recherche axée sur les données. Un point fort est l'application personnalisée au domaine médical, qui a nécessité l'adaptation des notions et des objectifs de l'argumentation pour convenir à ce domaine complexe. La thèse améliore notre compréhension théorique de la modélisation de la qualité et apporte des avancées pratiques dans l'extraction d'arguments. En reliant les idées entre les domaines, elle ouvre la voie à de futures recherches interdisciplinaires à l'intersection de l'argumentation, de l'apprentissage automatique et de disciplines spécialisées telles que les soins de santé.

**Mots clés:** Traitement Automatique du Langage Naturel, Extraction d'Information, Qualité de l'Argumentation, Fouille d'Arguments, l'IA au Service de la Médecine



## *Abstract*

The automated analysis of argumentation has garnered significant interest in recent years, as computational methods stand to enhance discourse quality across domains. This is especially pertinent in complex fields like healthcare, where sound reasoning bears direct impacts on human lives. The work presented in this thesis advances the state-of-the-art in argument mining and quality assessment, crafted to the intricacies of the medical domain. The thesis makes four main contributions: (1) Development and application of argument mining techniques, including analysis of their use in various domains and contributions to COVID-19 research. (2) Argumentation quality assessment methods, including annotation of a new dataset of 402 student essays with quality dimensions like cogency, rhetoric, and reasonableness. Innovative neural architectures combining textual features and graph embeddings are shown to aptly classify these facets, obtaining .78 F1, .89 F1, and 0.54 F1 respectively. (3) Identification of potential premises in the medical domain by automatically analyzing symptoms from 314 clinical cases and aligning them with external knowledge sources such as the Human Phenotype Ontology (HPO) using contextual embeddings (.53 accuracy). (4) Development of a transparent prevalence function to rank the explanatory power of the identified premises, leveraging statistics like abnormality and uniqueness from the knowledge base. This thesis makes significant contributions to the fields of argument mining and quality assessment through the development of novel techniques and resources. The proposed methods push the boundaries of automatic argument analysis, while the specially crafted datasets provide new opportunities for data-driven research. A major highlight is the tailored application to the medical domain, which required adapting argumentation notions and objectives to suit this complex field. The thesis enhances our theoretical understanding of quality modelling and delivers practical advancements in argument mining. By connecting insights across domains, it paves the way for future interdisciplinary research at the intersection of argumentation, machine learning, and specialized disciplines like healthcare.

**Keywords:** Natural Language Processing, Information Extraction, Argument Quality, Argument Mining, AI for Medicine





## *Acknowledgements*

It is hard to believe that this journey is coming to an end. Time has flown by, but I have learned so much during my Ph.D. I started out as an enthusiastic, yet naive student, but thanks to my supervisors, Serena and Elena, I have become a professional researcher. Without them, none of this would have been possible. I am infinitely grateful to them, not only for their guidance and endless opportunities to grow my career but also, and equally important, for their patience, empathy and hours and hours dedicated to helping me during my path.

I would also like to express my deepest and most sincere thanks to my family. My parents and sister have supported me every step of the way. They encouraged me to chase my dreams, even if that meant being apart for many many months a year. Nonetheless, I have always been able to feel their unconditional love and support, a safety net that always felt so secure and helped me to explore new horizons.

I want to thank Fabien for creating and nurturing a motivating and friendly team. I also want to thank the past and current members of our team, WIMMICS and SPARKS, for the coffee breaks and meals we shared almost every day. In particular, I want to thank my friends Amine, Ali, Anna, Iliana, Lucie and Nicholas, who always listened to my ideas, complaints, doubts, and issues. They helped me stay on track during this journey, and we were always able to laugh at our problems together. I hope we can continue to do so in the future.

I am grateful to my friends Theo, Milton, Magali, Marianela, Ignacio, Agustin, Ramiro, Federico, Rocio and Eugenia, who were always present for me when I needed to blow off some stress, meditate on why I was doing this and have endless talks about life, games, music and everything that makes us happy. I consider myself very lucky to have many supportive friends in my life.

Finally, I want to express my immense gratitude to Paula. I could not have done this without her. She has been my rock and my unconditional support for many years during my PhD. Her unwavering encouragement and belief in me gave me the strength to finish. The deep support and love she has for me is impossible to express and thank in a paragraph. Just to give an example, when a deadline was approaching she allowed me (on her own will) to fully focus on work, taking care of basic needs like doing groceries, cooking and so on. She helped me mentally and physically, making this journey not only easier but enjoyable as well. For that and infinite other things, I will be forever grateful for having her in my life. Thank you for everything, Paula.

**FUNDING:** This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. This work was supported by the CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002.



# Contents

<b>Résumé</b>	<b>v</b>
<b>Abstract</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>List of Published Papers</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Research Questions . . . . .	3
1.2.1 General Argument Quality Assessment . . . . .	3
1.2.2 Argument Quality Assessment in Healthcare . . . . .	4
1.3 Structure . . . . .	5
<b>2 Background</b>	<b>7</b>
2.1 Argument Mining . . . . .	7
2.2 Argument Quality Assessment . . . . .	9
2.3 Natural Language Processing for the Medical Domain . . . . .	14
2.3.1 Evidence-Based Medicine . . . . .	14
2.3.2 External Medical Knowledge for Quality Assessment . . . . .	15
2.4 Natural Language Representations . . . . .	16
2.4.1 Context-free Representations . . . . .	16
2.4.2 Contextualized Representations . . . . .	19
<b>3 Argumentation Quality</b>	<b>23</b>
3.1 Related Work . . . . .	24
3.2 Quality dimensions of persuasive essays . . . . .	25
3.3 Automatic assessment of argumentation . . . . .	31
3.3.1 Cogency and rhetoric scoring assessment . . . . .	32
3.3.2 Reasonableness scoring assessment . . . . .	34
3.4 Evaluation . . . . .	35
3.5 Concluding remarks . . . . .	40
<b>4 Integrating and Assessing External Knowledge in Medical Text</b>	<b>41</b>
4.1 Introduction . . . . .	42
4.2 Related Work . . . . .	44

4.3	Extracting and Aligning Clinical Information . . . . .	48
4.3.1	Dataset . . . . .	48
4.3.2	Proposed framework . . . . .	50
4.3.3	Evaluation . . . . .	51
4.4	Assessing Explanatory Power . . . . .	55
4.4.1	Assessing Reasons used in Explanations . . . . .	57
4.4.2	Data Preprocessing . . . . .	57
4.4.3	Identification and Alignment of Potential Causes . . . . .	58
4.4.4	Prevalence Function . . . . .	59
4.4.5	Reason Alignment via Sentence Matching . . . . .	60
4.4.6	Template-Based Explanation Generation . . . . .	62
4.4.7	Evaluation . . . . .	63
4.5	Natural Language Explanation Generation . . . . .	65
4.6	Concluding Remarks . . . . .	69
<b>5</b>	<b>Argument Mining on Clinical Trials</b>	<b>71</b>
5.1	Covid-on-the-Web project . . . . .	72
5.1.1	The Covid-on-the-Web RDF Dataset . . . . .	73
5.2	Argument Mining on Clinical Trials: Computing the Effect-on-Outcome . . . . .	76
5.2.1	Experimental Setup . . . . .	77
5.2.2	Results and Discussion . . . . .	78
5.3	ACTA 2.0 . . . . .	80
5.3.1	Main Functionalities . . . . .	81
5.4	Open Challenges . . . . .	84
<b>6</b>	<b>Conclusion and Future Perspectives</b>	<b>85</b>
	<b>Bibliography</b>	<b>91</b>

# List of Figures

2.1	An illustrative example of an Argumentation Mining pipeline. Figure drawn from [22]. . . . .	9
2.2	The proposed taxonomy of argumentation quality as well as the mapping of existing assessment approaches to the covered quality dimensions. Arrows show main dependencies between the dimensions. Figure drawn from [39]. . . . .	13
2.3	The transformer model architecture. Figure drawn from [63]. . . . .	20
3.1	Example of an argument graph of a persuasive essay [81]. . . . .	30
3.2	Overview of our natural language argumentation quality prediction model. . . . .	34
4.1	Overview of our full pipeline for symptom prediction and alignment, and NL explanation generation module. . . . .	51
4.2	Overview of our approach for the automatic assessment of explanation’s reasons. . . . .	58
5.1	Illustration of the Covid-on-the-Web [170] pipeline, its services and applications. . . . .	75
5.2	Illustration of the full Argument Mining pipeline with the outcome analysis extension. . . . .	77
5.3	Confusion matrix of the predictions on the test set of the outcome classification. . . . .	79
5.4	Multiple screenshots to illustrate the different functionalities of ACTA and the visualization of the argument graph returned to the user. . . .	82



# List of Tables

3.1	Analytic Scoring Rubric to assess Cogency [19]. . . . .	27
3.2	Analytic Scoring Rubric for assessing Reasonableness Counterargument [19]. . . . .	28
3.3	Analytic Scoring Rubric for assessing Reasonableness Rebuttal [19]. . . . .	29
3.4	Statistics of the dataset, reporting on the percentage of Cogency and Reasonableness for each score. . . . .	31
3.5	Statistics of the dataset, reporting on the percentage and type of Rhetorical arguments. . . . .	31
3.6	Results for the Cogency score of the 3-class sequence tagging task are given in weighted F1 (f1) and macro F1 (F1). . . . .	37
3.7	Results of the Argumentation Rhetoric sequence tagging task training an SVM model (weighted F1 (f1) and macro F1 (F1)). . . . .	37
4.1	Statistics of the MEDQA-USMLE-Symp dataset. . . . .	49
4.2	Results for entity recognition in macro multi-class precision, recall, and F1-score. . . . .	54
4.3	Results for entity recognition using our best performing model (SciBERT uncased) in P, R, and F1-score. . . . .	54
4.4	Results for DASH and our symptom alignment method using different embeddings with and without context (accuracy score). . . . .	55
4.5	Results for named entity-based matching in macro multi-class precision, recall, and F1-score. . . . .	64
5.1	Statistics of the Outcome dataset, showing the numbers of <i>Improved</i> , <i>Increased</i> , <i>Decreased</i> , <i>NoDifference</i> and <i>NoOccurrence</i> classes independent of the disease-based subsets. . . . .	77
5.2	Results for the outcome analysis pipeline, given in overall macro $F_1$ and label-wise binary $F_1$ -score. . . . .	78





# List of Abbreviations

<b>ACTA</b>	<b>Argumentative Clinical Trial Analysis</b>
<b>AI</b>	<b>Artificial Intelligence</b>
<b>AM</b>	<b>Argument Mining</b>
<b>AMO</b>	<b>Argument Model Ontology</b>
<b>API</b>	<b>Application Programming Interface</b>
<b>BERT</b>	<b>Bidirectional Encoder Representations from Transformers</b>
<b>BOW</b>	<b>Bag-of-Words</b>
<b>CC</b>	<b>Clinical Case</b>
<b>CORD-19</b>	<b>COVID-19 Open Research Dataset</b>
<b>CRF</b>	<b>Conditional Random Field</b>
<b>CUI</b>	<b>Concept Unique Identifiers</b>
<b>EBM</b>	<b>Evidence-based Medicine</b>
<b>ELMo</b>	<b>Embeddings from Language Model</b>
<b>GRU</b>	<b>Gated Recurrent Unit</b>
<b>HPO</b>	<b>Human Phenotype Ontology</b>
<b>IAA</b>	<b>Inter-Annotator Agreement</b>
<b>IOB</b>	<b>Inside–Outside–Beginning</b>
<b>JSON</b>	<b>JavaScript Object Notation</b>
<b>KB</b>	<b>Knowledge Base</b>
<b>LM</b>	<b>Language Model</b>
<b>LLM</b>	<b>Large Language Model</b>
<b>LOD</b>	<b>Linked Open Data</b>
<b>LSTM</b>	<b>Long Short-Term Memory</b>
<b>ML</b>	<b>Machine Learning</b>
<b>MLM</b>	<b>Masked Language Modeling</b>
<b>MSE</b>	<b>Mean Squared Error</b>
<b>NE</b>	<b>Named Entity</b>
<b>NER</b>	<b>Named Entity Recognition</b>
<b>NLI</b>	<b>Natural Language Inference</b>
<b>NLP</b>	<b>Natural Language Processing</b>
<b>NLU</b>	<b>Natural Language Understanding</b>
<b>NN</b>	<b>Neural Network</b>
<b>OOV</b>	<b>Out-of-Vocabulary</b>
<b>PICO</b>	<b>Population Intervention Comparison Outcome</b>
<b>PMID</b>	<b>PubMed Identifier</b>

<b>RCT</b>	<b>R</b> andomized <b>C</b> ontrolled <b>T</b> rial
<b>RDF</b>	<b>R</b> esource <b>D</b> escription <b>F</b> ramework
<b>RNN</b>	<b>R</b> ecurrent <b>N</b> eural <b>N</b> etwork
<b>RQ</b>	<b>R</b> esearch <b>Q</b> uestion
<b>SBERT</b>	<b>S</b> entence- <b>B</b> ERT
<b>SNLI</b>	<b>S</b> tanford <b>N</b> atural <b>L</b> anguage <b>I</b> nterference <b>C</b> orpus
<b>SOTA</b>	<b>S</b> tate- <b>o</b> f- <b>t</b> he- <b>A</b> rt
<b>SPARQL</b>	<b>S</b> PARQL <b>P</b> rotocol and <b>R</b> DF <b>Q</b> uery <b>L</b> anguage
<b>SVM</b>	<b>S</b> upport <b>V</b> ector <b>M</b> achine
<b>T5</b>	<b>T</b> he <b>T</b> ext- <b>T</b> o- <b>T</b> ext <b>T</b> ransformer
<b>tf-idf</b>	<b>T</b> erm <b>F</b> requency- <b>I</b> nverse <b>D</b> ocument <b>F</b> requency
<b>UMLS</b>	<b>U</b> nified <b>M</b> edical <b>L</b> anguage <b>S</b> ystem
<b>U.S</b>	<b>U</b> nited <b>S</b> tates
<b>URI</b>	<b>U</b> niform <b>R</b> esource <b>I</b> dentifier
<b>XAI</b>	<b>E</b> Xplainable <b>A</b> rtificial <b>I</b> ntelligence

# List of Published Papers

**Santiago Marro**, Benjamin Molinet, Elena Cabrio, Serena Villata. "Natural Language Explanatory Arguments for Correct and Incorrect Diagnoses of Clinical Cases", In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence 2023: ICAART 2023*, vol. 1, pages 438-449, SciTePress.

**Santiago Marro**, Elena Cabrio and Serena Villata. "Graph Embeddings for Argumentation Quality Assessment", In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4154–4164, Association for Computational Linguistics.

**Santiago Marro**, Elena Cabrio and Serena Villata. "Argumentation quality assessment: an argument mining approach", European Conference on Argumentation (ECA) 2022, College Publications.

Benjamin Molinet, **Santiago Marro**, Elena Cabrio, Serena Villata, and Tobias Mayer. "ACTA 2.0: A Modular Architecture for Multi-Layer Argumentative Analysis of Clinical Trials." In *IJCAI 2022-Thirty-First International Joint Conference on Artificial Intelligence: IJCAI-ECAI 2022*. Demo Paper on the Demo Track. Pages 5940-5943

Damian Furman, **Santiago Marro**, Cristian Cardellino, Diana Popa, Laura Alonso Alemany. "You can simply rely on communities for a robust characterization of stances", In the *Florida Artificial Intelligence Research Society proceedings FLAIRS 2021*, vol 34, AAAI Press

Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, **Santiago Marro**, Tobias Mayer, Mathieu Simon, Serena Villata, Marco Wincker, "Covid-on-the-Web: Graphe de Connaissances et Services pour faire Progresser la Recherche sur la COVID-19", In *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA'21) 2021*, page 113. **Best Paper Award** obtained.

Tobias Mayer, **Santiago Marro**, Elena Cabrio and Serena Villata, "Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials". In: *Artificial Intelligence in Medicine 2021*. Vol. 118, page 102098, Elsevier.

Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio,

Olivier Corby, Raphaël Gazzotti, Alain Giboin, **Santiago Marro**, Tobias Mayer, Mathieu Simon, Serena Villata and Marco Winckler, "Covid-on-the-web: Knowledge graph and services to advance covid-19 research". In *Proceedings of the 19th International Semantic Web Conference (ISWC)*, 2020, The Semantic Web – ISWC 2020 book, pages 294-310, Springer International Publishing

Tobias Mayer, **Santiago Marro**, Elena Cabrio, and Serena Villata, "Generating Adversarial Examples for Topic-Dependent Argument Classification". In *Proceedings of the 8th International Conference on Computational Models of Argument (COMMA)*, 2020, pages 33-44, IOS Press.

### **Papers under review**

**Santiago Marro**, Theo Alkibiades Collias, Elena Cabrio, Serena Villata. "On the Automatic Assessment of Natural Language Expert Explanations in Medicine", In the *Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2023)*

# Chapter 1

## Introduction

*In this chapter, the motivation behind the research presented in this thesis is discussed. The need to automatically evaluate the quality of argumentative text for educational purposes, particularly in the medical field, is emphasized. Furthermore, the inclusion of external knowledge is deemed necessary to model the expert's reasoning in tackling this challenge. This is especially important when used in conjunction with established frameworks for evidence categorization, such as medical-named entities.*

### 1.1 Background and Motivation

Argumentation is fundamental to many facets of society, enabling the justification of claims and decisions in domains like law, politics, and medicine [1]. However, not all arguments are equally strong or persuasive. Different dimensions have been proposed to assess argumentative quality such as Cogency, Reasonableness and Effectiveness [2]. Moreover, evaluating argument quality is an hard task. For instance, assessing logical cogency requires analyzing if premises are acceptable, relevant, and collectively sufficient, which demands close reading and logical reasoning, or determining rhetorical persuasiveness involves subjective judgments of emotional appeal, style, and credibility that vary across audiences [3].

In recent years, the proliferation of argumentative text requires the definition of automatic methods to assess the quality of such text. This task is particularly relevant in the educational context, where students need to produce high quality text, e.g., student persuasive essays. Specialized domains pose further challenges, where a precise domain expertise is required for informed quality assessment, e.g., clinical knowledge is needed to evaluate medical arguments. Moreover, quality metrics designed for general arguments may not capture domain-specific notions. Furthermore, quality assessment suffers from subjectivity and inconsistency. While techniques have emerged for computational argument mining and assessment [4, 5], gaps remain in adapting these models to the real-world settings.

Recent years have seen a growing interest in developing AI-based healthcare systems that can support and simplify everyday tasks for clinicians. These systems handle various types of data, including medical images, biometrics, and textual documents such as electronic health records and clinical guidelines. Some examples of such solutions include Evidence-based reasoning for decision-making [6–8] and automated medical entity extraction [9–13] where the entities range from drug-disease interactions to the identification of diseases, genes, and molecular entities such as protein, DNA, RNA. The goal is to extract the necessary information from unstructured textual documents and present it in a structured manner, making it easy for clinicians to analyze [14]. In this context, the assessment of argumentation quality has been tackled as the assessment of medical evidence, with the emergence of Evidence-Based Medicine (EBM). More precisely, the purpose of EBM is to improve the quality of medical evidence by establishing systematic evaluation standards and reducing bias in the reports. EBM emphasizes the critical appraisal and judicious use of evidence, combining high-quality evidence with the clinician’s individual clinical experience and the patient’s values to achieve the best possible outcome [15]. Moreover, EBM should help clinicians keep up to date with the latest research findings and incorporate them into their everyday decision-making. The focus has shifted to identifying the best available evidence empirically to ensure its quality. When diagnosing a patient, physicians must make decisions based on the available evidence, comparing evidence from trials or guidelines with the patient’s individual circumstances. They must determine if the evidence matches the patient’s unique characteristics and whether the potential costs and benefits are reasonable. EBM should provide the scientific framework necessary for optimal healthcare, from the systematic evaluation of evidence to the facilitation of the decision-making process [15–18].

However, to properly assess an argument in the medical domain, the evidence must be evaluated with respect to established medical knowledge. The reasoning must be supported by and integrated with the practitioner’s expertise [15]. This highlights the need for argumentational quality assessment methods that model clinical knowledge to judge the acceptability of claims and the coherence of the reasoning. Moreover, residents must learn how to construct cogent rationales that align with clinical knowledge as practitioners do. Hence, transparent quality assessment methods are needed in such a way that each step taken in evaluating student arguments is justified, highlighting flaws in reasoning and integrating domain expertise for teaching purposes. The goal is to develop argumentation quality assessment techniques that balance domain-specific knowledge with interpretability, to effectively evaluate the quality of medical argumentation, providing pedagogical insights to students. Advances in computational argumentation quality assessment are required to handle the complexity of clinical reasoning.

*Hence, the goal of this PhD thesis is two-fold: (i) defining novel methods to automatically assess the quality of textual arguments, with a focus on educational use cases such as student persuasive essays, and (ii) enhance these methods with knowledge-based algorithms to enrich and assess the quality of clinical evidence in the medical domain.*

## 1.2 Research Questions

We established a road map comprising multiple stages to execute this project. Each stage was further divided into a research question (RQ) that addressed different aspects of the project:

### 1.2.1 General Argument Quality Assessment

Evaluating the quality of argumentation is critical yet challenging, as high-quality arguments should demonstrate cogency, reasonableness, rhetorical effectiveness, and other key attributes. However, manually assessing large volumes of argumentation along these dimensions is infeasible. My first research line aims to develop computational methods that can automatically assess argument quality in a multi-dimensional manner, and it corresponds to the first two research questions I answered in this thesis.

**RQ1** : *How can we model the multi-dimensional notions of quality to capture key aspects like cogency, rhetoric, and reasonableness?*

I addressed this research question by first defining three prominent quality dimensions for natural language argumentation - cogency, rhetoric, and reasonableness - based on the literature in argumentation theory and social science scoring rubrics for persuasive writing [2, 19, 20]. I then annotated these dimensions on a corpus of 402 student persuasive essays to create a novel annotated resource [20].

**RQ2** : *Can integrating textual features with graph embeddings and emotion detection improve the assessment of argumentation quality dimensions like cogency, rhetoric and reasonableness?*

To computationally assess these annotated quality dimensions, I proposed a novel neural architecture that exploits the inherent graph structure of argumentation frameworks, in combination with textual features. Specifically, I showed that augmenting contextualized text embeddings (e.g., BERT [21]) with graph embeddings and emotion detection models significantly improves the predictive performance across the quality dimensions, increasing macro F1 scores by 5-10 percentage points compared to baselines relying solely on textual features [20].



By encoding textual semantics along with topological structure and rhetoric cues, I demonstrated a more reliable modelling of quality attributes like cogency and reasonableness. These results highlight the benefits of an integrated approach combining graphical and linguistic information for argument quality prediction.

Through this investigation, I established a general framework for automatically evaluating key dimensions of argumentation quality on a linguistic dataset for educational purposes. However, assessing real-world argumentation in specialized domains poses additional challenges I sought to address.

#### Related Publications:

- **Santiago Marro**, Elena Cabrio and Serena Villata. "Graph Embeddings for Argumentation Quality Assessment", In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4154–4164, Association for Computational Linguistics.
- **Santiago Marro**, Elena Cabrio and Serena Villata. "Argumentation quality assessment: an argument mining approach", European Conference on Argumentation (ECA) 2022, College Publications.

### 1.2.2 Argument Quality Assessment in Healthcare

While in the first research line of this thesis I proposed novel methods for argument quality modelling in general domains, clinical argumentation poses unique challenges needing specialized methods. Medical arguments demand precise, verified knowledge to justify claims, differently from informal domains where rhetorical flair or emotional appeals may suffice. I tailored argumentation quality assessment to handle the intricacies of the medical field, leading me to answer two further research questions.

**RQ3** : *Which kind of adaptations are required to assess clinical argumentation given its specialized nature?*

To answer this research question, it is worth noticing that dimensions like cogency, rhetoric, and reasonableness capture important attributes, but lack the precision and validation required in medicine. Soundness of premises is paramount, requiring external verification rather than internal sufficiency. To tackle these challenging issues, I proposed new methods to enhance existing quality models with finer-grained validation of medical knowledge. This raises an additional question about integrating external medical evidence.

**RQ4** : *How can external knowledge be integrated to enrich the analysis of clinical decision making and their explanation?*

I explored harnessing medical knowledge bases to extract salient clinical entities from text and map them to established medical ontologies. By semantically enriching arguments with the aligned ontology terms, I enabled assessing the veracity and relevance of the evidence used in the explanations of clinical diagnoses. The specialized pipeline I proposed extracts and aligns medical concepts to improve quality evaluation with validated domain knowledge. The investigation of clinical quality assessment pushed boundaries in adapting argument mining and quality models to leverage external domain information critical for the healthcare domain. The work opened promising directions at the intersection of quality modelling and knowledge-rich reasoning in sensitive applications.

#### Related Publications:

- **Santiago Marro**, Benjamin Molinet, Elena Cabrio, Serena Villata. "Natural Language Explanatory Arguments for Correct and Incorrect Diagnoses of Clinical Cases", *In Proceedings of the 15th International Conference on Agents and Artificial Intelligence 2023: ICAART 2023*, vol. 1, pages 438-449, SciTePress.
- **Santiago Marro**, Theo Alkibiades Collias, Elena Cabrio, Serena Villata. "On the Automatic Assessment of Natural Language Expert Explanations in Medicine", *In the Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2023)* (under review)

## 1.3 Structure

The thesis is organized as follows:

**Chapter 2** describes the preliminaries, which are used throughout the thesis. It provides insights into the context and practices of the applied domain, i.e., evidence-based medicine. Further, the main concepts and open challenges in the research fields of Argument Mining and argument quality are presented.

**Chapter 3** presents the argument quality assessment framework for persuasive essays. First, annotation guidelines were developed and used to annotate 402 persuasive essays with three quality dimensions: Cogency, Reasonableness, and Rhetorical strategy. The framework addresses the automatic assessment of these dimensions. Methods for Cogency and Rhetorical strategy classification include embedding-based SVMs, Random Forests, and fine-tuned transformer models. Reasonableness assessment is addressed in two ways: by modelling it based only on textual and graph features, and with a transparent deterministic algorithm leveraging the graph structure and the cogency of arguments. Results are discussed along with an in-depth error analysis.

**Chapter 4** introduces novel methods focusing on argumentation quality in medicine. Specifically, I present methods to automatically annotate medical information from clinical cases and, using external knowledge such as the Human Phenotype Ontology, to assess potential premises explaining a patient's diagnosis. The framework generates template-based explanations using the best-assessed premises, to improve student explanations.

**Chapter 5** demonstrates the applications of the proposed approaches in the medical domain, spanning from the Covid-on-the-Web project developed to address an argumentative analysis of the clinical articles about Covid-19 to the study of the effect of interventions on the outcomes for the AbsRCT dataset [22]. A Proof-of-Concept online system, ACTA 2.0, provides argumentative analysis of medical articles to support clinicians decision-making in real-time, integrating the effect on outcomes analysis and a modular pipeline implementation allowing researchers to exploit each one of the steps in the system separately.

**Chapter 6** concludes the thesis by summarizing the main contributions. Furthermore, perspectives for future applications and further research directions are proposed, as well as potential plans for improvements.

## Chapter 2

# Background

*In this chapter, the preliminaries used throughout the thesis are introduced. We provide insights into the concepts and challenges in the research field of Argument Mining with a focus on argumentation quality assessment. Further, given that this thesis investigates the application of Argument Mining to the medical domain, this chapter outlines the practices of Evidence-Based Medicine, presenting the concepts and challenges in Argument Mining and Quality Assessment, and the Natural Language Processing methods employed.*

This chapter provides the key background concepts and methodologies that enable the technical contributions presented later in the thesis. First, Section 2.1 introduces the field of Argument Mining, including core tasks like argument component extraction and relation prediction. Argument Mining provides a foundation for assessing argument quality, which is described next in Section 2.2. This section details the motivation, tasks, and challenges involved in evaluating key qualities like cogency and reasonableness for natural language arguments. Section 2.2 emphasizes the need for specialized techniques tailored to assess the quality of clinical arguments in the medical domain. To provide context on this application area, Section 2.3 expands on two relevant aspects of healthcare. An overview of Evidence-Based Medicine is given, where argument mining and quality assessment can assist doctors in analyzing clinical trial reports. Furthermore, the use of external knowledge bases to enrich the assessment of medical arguments is discussed. Finally, Section 2.4 surveys the natural language processing methods leveraged in this thesis to computationally represent text. Early context-free models like bag-of-words are presented, along with recent advanced contextualized models.

### 2.1 Argument Mining

Argumentation is a well-established interdisciplinary field at the intersection of natural language processing, computational linguistics, and artificial intelligence. Argumentation is the process by which arguments are constructed, compared, evaluated in several respects and judged in order to establish whether any of them is

warranted [23, 24]. It is an effective approach for solving various theoretical and practical problems, like explaining and justifying the decision-making outcomes and reasoning under inconsistent and incomplete information. Roughly, each argument is a set of premises or assumptions that, together with a claim, is obtained by a reasoning process. The overall goal of argumentation is to increase or decrease the acceptability of claims by supporting or attacking them with new arguments. However, argument-based decision-making requires structured input. In most real-world contexts, argumentative texts are presented in unstructured natural language without explicit argument components, necessitating the development of computational methods to automatically extract structured arguments. One of the latest advances in artificial argumentation [1], which tackles the aforementioned problem, is the so-called *Argument(ation) Mining (AM)* [25–28].

The goal of argumentation mining is to automatically identify argumentative structures within texts by detecting claims, premises, and the relationships between them [25]. This capability has numerous potential applications, including analyzing persuasiveness in essays, understanding reasoning in legal documents, political debates, countering disinformation and analyzing clinical trials.

Several pioneering works introduced the problem of mining arguments from text, though they did not initially gain much traction in the NLP community. One of the earliest is argumentative zoning [29], where sentences are classified by their rhetorical role in a scientific paper (e.g. background, objectives). While not extracting full argument structure, this classification paved the way for later AM approaches [30]. Other early work includes detecting argument components in legal text [31, 32]. However, these approaches were limited by the NLP techniques available at the time. As methods for computationally processing natural language advanced, enabling more complex tasks like AM, interest increased [26]. AM itself requires deep natural language understanding and is closely related to natural language inference, leading initial techniques to draw inspiration from textual entailment [33, 34]. With machine learning and NLP now significantly more mature, researchers can effectively develop novel AM methods.

Most work conceptualizes argumentation mining as consisting of two main tasks:

1. **Argument extraction:** Identifying argument components like claims, premises, and evidence within a text. This may be further divided into detection and segmentation subtasks.
2. **Relation prediction:** Predicting the relationships between extracted arguments, like support, attack, or entailment. This enables constructing full argument graphs.

Figure 2.1 showcases an example of an argumentation mining pipeline in the context of clinical trials.

Within these tasks, supervised machine learning techniques predominate. Support vector machines, recurrent neural networks, conditional random fields, and

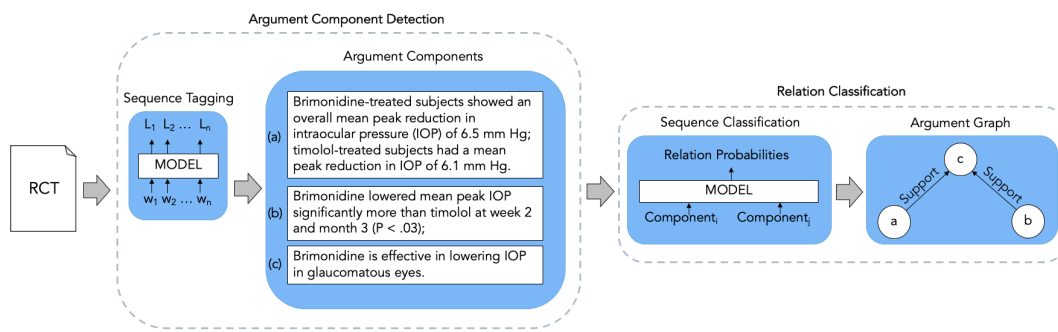


FIGURE 2.1: An illustrative example of an Argumentation Mining pipeline. Figure drawn from [22].

other algorithms are commonly applied [26]. As with any supervised approach, annotated corpora are critical for training and evaluation. Although diverse datasets have been constructed across domains like student essays, legal decisions, political speeches, and social media, their heterogeneity presents challenges for collective progress [35]

Beyond the core goals of detecting argument components and predicting relationships, numerous nuanced subtasks have emerged over time that aims to enrich the extracted structures with additional informative features advantageous for various application scenarios. The extracted argument structures can be integrated with formal argumentation models to enable advanced tasks like identifying fallacies, assessing argument quality, evaluating student essays, clustering arguments, determining argument relevance, detecting rhetorical figures, and classifying fine-grained evidence types.

## 2.2 Argument Quality Assessment

The assessment of argument quality is a critical aspect of computational argumentation. While simple acceptability calculations can determine the justification status of abstract arguments [36], they only represent a (basic) part of the complex assessment tasks required in argumentative processes in many everyday life applications and contexts, e.g., in medicine and education. Assessing natural language argumentation involves examining logical, rhetorical, and dialectical dimensions across different levels of analysis [2].

Consider the following argument against abortion:

**Example 2.2.1** *"The fetus has a right to live, so abortion is morally wrong."*

Although the conclusion is implicit, it seems logically sound. However, some people may reject the premise, especially if they prioritize women's rights. Or they may doubt this is the most relevant argument in the abortion debate. This example reveals three key challenges in assessing argument quality:

1. Quality is evaluated across different granularities, from individual units to full texts.
2. Many quality dimensions are subjective, depending on people's preconceived opinions.
3. Overall quality seems difficult to measure due to complex interactions between dimensions.

Several theoretical frameworks have examined the characteristics of strong, high-quality arguments. One influential work is Blair's argumentation theory [2]. Blair proposes three overarching qualities for assessing argumentation:

**Cogency** This refers to the logic and structure of an argument. A cogent argument has premises that are acceptable, relevant, and jointly sufficient to support the conclusion [2]. Acceptability means the premises are worthy of belief by the audience. Relevance indicates the premises contribute toward the conclusion. Sufficiency means enough evidence is provided to rationally justify the conclusion.

Let us explore two examples extracted from Johnson and Blair [37] of what are considered good and bad arguments in terms of cogency. Example 2.2.2 deals with highway speed limits. In the seventies, the speed limit on interstate highways was reduced from 70 to 55 miles per hour. In an article in *En Route* magazine, Len Coates objected as follows:

**Example 2.2.2** *Yes, it is true that the 55 mph saves lives. The National Highway Traffic Safety Administration estimates that 4,500 lives have been saved by the 55 mph limit. But surely there are more cost-efficient ways of saving lives . . . such as equipping every house with a smoke detector (that would cost \$50,000 to \$80,000 per life) or putting more dialysis machines in hospitals (\$30,000 per life).*

**Example 2.2.3.** This is an excerpt from Josiah Thompson's book *Six Seconds in Dallas* (New York: Bernard Geis, 1967) about the assassination of President John F. Kennedy. Thompson is discussing the question, Where did the first bullet go (p. 39).

**Example 2.2.3** *The testimony of Secret Service Agent Roy Kellerman adds weight to the theory that the first bullet only lodged in the President's back. Seated in the right front seat of the presidential limousine, Kellerman heard Kennedy yell, "My God! I'm hit" just after the first shot. . . . Since the projectile that caused the throat wound ripped his windpipe in passing, it seems unlikely that the President could have spoken after receiving the throat wound.*

As Johnson and Blair [37] point out, the argument presented in Example 2.2.2 is flawed. Although it may be factual that installing smoke detectors in every house could save lives, it is not relevant to the issue of saving lives on the highway, which is the focus of the argument. The installation of smoke detectors is not impeded by reducing speed limits.

On the other hand, Example 2.2.3 is a fairly strong argument. If the first bullet pierced Kennedy's throat (as some allege), then Kellerman could not have heard what he said he heard. Hence his testimony "adds weight to the theory that the first bullet only lodged in the President's back." Thompson's conclusion is presented in a qualified way ("adds weight," "seems unlikely"), and he presents contrary evidence later in the book (no one else heard what Kellerman heard the President say). But if the facts are as recorded in Example 2.2.3, they provide fairly compelling evidence that the first bullet did not pierce Kennedy's throat but lodged in his back [37].

**Effectiveness** - This relates to the persuasive rhetorical style and arrangement of an argument. An effective argument uses language, reasoning, and emotional appeals tailored to the audience and situation in order to achieve adherence to its conclusion [2]. Arrangement and clarity are important factors.

Given the two following examples:

**Example 2.2.4** *Regular exercise is beneficial for maintaining good health. Numerous studies have shown that individuals who engage in regular physical activity have a lower risk of developing chronic diseases such as heart disease, diabetes, and obesity. Additionally, exercise has been linked to improved mental health and well-being.*

The effectiveness of the argument presented in Example 2.2.4 lies in its use of clear language, logical reasoning, and credible sources to support the claim. It is designed to be persuasive and easy to follow, tailored to the audience's knowledge of health and well-being. Additionally, it presents evidence from trustworthy sources, making it even more convincing.

**Example 2.2.5** *Drinking coffee is bad for your health. My friend John drinks coffee daily and was recently diagnosed with high blood pressure.*

Example 2.2.5 presents a weak argument that relies on anecdotal evidence and a hasty generalization fallacy. The argument is based on a single example that may not be representative of the broader population, and the evidence provided is not from a credible source nor does it address counterarguments or other factors that could contribute to high blood pressure. The argument's lack of clarity and organization makes it less persuasive and more challenging to follow its conclusion.

**Reasonableness** - This considers how an argument contributes to critically resolving an issue through dialogue. A reasonable argument provides information acceptable to the audience that helps arrive at a mutually satisfactory conclusion [2, 38]. Reasonableness indicates the argument appropriately moves the discussion forward.

**Example 2.2.6** *Implementing a recycling program in our community will reduce waste and promote environmental awareness. Studies have shown that communities with recycling programs experience a significant reduction in waste sent to landfills and an increase in resource*



conservation. Additionally, recycling programs have been linked to increased environmental awareness among residents.

**Example 2.2.7** *We should not implement a recycling program in our community because it is too expensive. My neighbour told me that their community's recycling program costs a lot of money and has not made a significant impact on waste reduction.*

The reasonableness of Example 2.2.6 can be justified as it discusses a pertinent matter and presents reliable facts that can be acknowledged by the readers. The argument follows a logical sequence, and the supporting evidence corroborates the assertion that introducing a recycling program would result in beneficial environmental outcomes. The discussion effectively presents a solution to the issue and encourages conversations around ecological accountability.

However, the argument made in Example 2.2.7 is flawed because it lacks sufficient data to support the claim that recycling programs are not cost-effective. It relies on anecdotal evidence and a single example, which may not be a reliable representation of other recycling programs or communities. The argument does not contribute to a constructive resolution of the issue, as it fails to offer any alternative solutions or acknowledge the potential benefits of recycling programs. Instead, it impedes productive discussion by not presenting strong evidence and failing to engage in critical dialogue.

Blair synthesizes various perspectives from informal logic and argumentation theory to unify these dimensions [2]. He details conceptual nuances differentiating between local, global, and dialectical aspects of quality. Blair's integration of logical, rhetorical, and dialectical factors provides a robust theoretical foundation for examining argumentation quality.

Computational approaches leverage this theoretical foundation. Wachsmuth et al. [39] derive a taxonomy organizing 15 fine-grained dimensions of argument quality. The taxonomy is illustrated in Figure 2.2. Their annotated corpus covers arguments from debate portals rated by experts along the taxonomy. Recent approaches in Argument(ation) Mining (AM) tackle specific argument qualities features, such as argument relevancy [40], convincing arguments [41] and overall argument quality [42]. Related work scores essay qualities like evidence [43] and organization [44].

Several datasets exist for studying argument quality, each capturing different notions of quality. The ArgQuality Corpus [39] contains 320 arguments annotated for 15 dimensions like cogency, effectiveness, and reasonableness. While comprehensive, the small scale limits use. Ng et al. [45] introduce a cross-domain corpus of 5,295 arguments labeled for cogency, reasonableness, effectiveness, and overall strength. However, the domains of online debates, QA forums, and reviews have limited diversity. Persing and Ng [46] collected student essay arguments rated for overall persuasiveness. Gretz et al. [47] crowdsourced arguments labeled by recommendability. While such corpora provide useful training data, they are restricted to

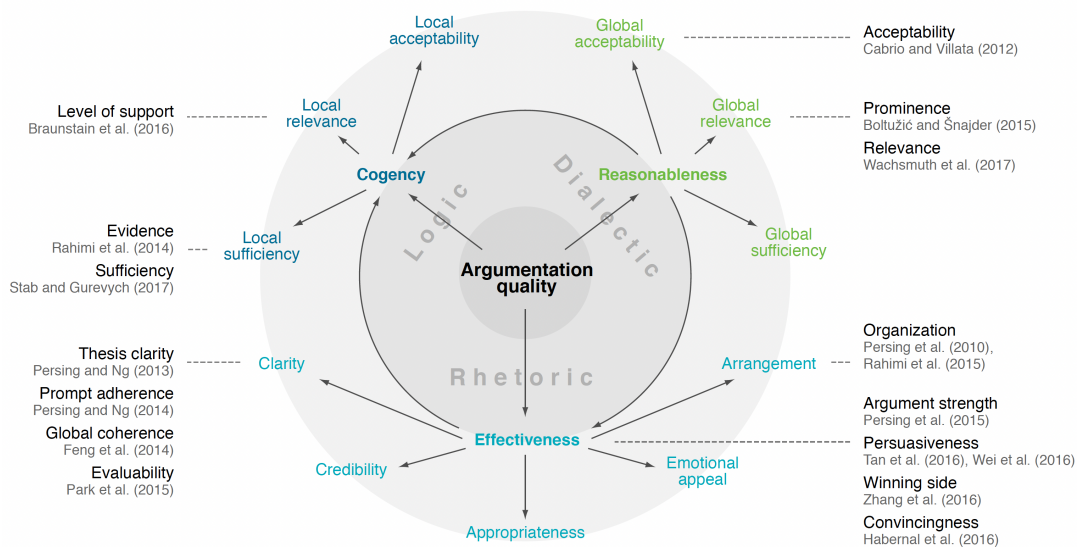


FIGURE 2.2: The proposed taxonomy of argumentation quality as well as the mapping of existing assessment approaches to the covered quality dimensions. Arrows show main dependencies between the dimensions. Figure drawn from [39].

certain genres, topics, and quality definitions. Notably, no quality assessment resources exist for the complex medical domain, where factors like clinical evidence and reasoning impact argument strength. The medical field's intricacies indicate that further research is needed to model quality, potentially requiring new specialized dimensions tailored to clinical arguments. Overall, existing resources exhibit a lack of diversity in terms of domains, topics, linguistic styles, and notions of quality.

Several promising medical applications could benefit from advances in assessing argument quality. However, major knowledge gaps exist in adapting argumentation quality assessment to the medical domain. This thesis addresses these gaps and contributes methods specialized for healthcare arguments. First, we propose computational techniques to leverage graphical argument structures, combined with textual features, to evaluate cogency and reasonableness - two key quality dimensions of clinical arguments. Second, we develop information extraction and external knowledge linking methods tailored to clinical cases, enabling inference for quality assessment. Third, we introduce a ranking and assessment framework that compares student explanations against high-quality premises extracted and inferred from cases. This supports argumentative writing education by suggesting improvements to student rationales. Overall, we investigate quality evaluation capabilities needed for medical applications through spanning graph-enhanced models, clinical information extraction, and assessment algorithms that leverage inferred knowledge. By targeting the complexities of the medical domain, we take significant steps toward quality-aware dialogue systems for education and social media.

## 2.3 Natural Language Processing for the Medical Domain

Evidence-based medicine (EBM) provides the clinical practice framework that motivates much of the thesis research. The first subsection defines EBM and describes how argumentation mining and quality assessment can assist core EBM processes like systematic reviews and appraising evidence. Realizing quality assessment innovations for EBM requires drawing on external domain knowledge. The second subsection introduces key external medical knowledge resources, specifically the Unified Medical Language System and the Human Phenotype Ontology. These structured vocabularies and ontologies enable grounding argumentation analysis in statistical, relational, and definitional data about clinical concepts. Together, the two subsections situate EBM as a driving application area and highlight the importance of external knowledge for enhancing quality assessment in the medical domain.

### 2.3.1 Evidence-Based Medicine

Evidence-based medicine (EBM) is an approach to clinical practice that emphasizes the use of high-quality scientific evidence to guide medical decision-making. The principles of EBM were first articulated in the 1990s as a way to shift medical practice away from reliance on expert opinion and pathophysiological rationales alone towards carefully evaluated evidence from clinical research studies and trials. [17]

The practice of EBM involves five key steps:

1. Assessment of the clinical problem,
2. Converting clinical problems into well-built questions,
3. Searching the literature for relevant evidence that can answer those questions,
4. Critically appraising the evidence for validity, impact, and applicability, and
5. Integrating the appraised evidence with clinical expertise and patient values to make decisions

Steps 2-4 highlight the areas most relevant to argumentation mining and quality assessment. To find high-quality evidence, EBM relies heavily on systematic reviews and meta-analyses of clinical trials. These provide a rigorous synthesis of all available evidence related to a focused clinical question. However, constructing systematic reviews requires meticulous analysis of many study reports to extract key information on methods, results, and conclusions. This process is time-consuming and difficult to scale.

Argumentation mining techniques can assist in automating parts of the review process, especially extracting argument components like claims and evidence from trial reports [48]. This can accelerate evidence synthesis and ensure all relevant information is considered. Further, argument mining enables the building of graphical

representations of the argument structure within trials, revealing how claims relate to supporting evidence.

Assessing the quality of clinical evidence is also critical in EBM. Not all evidence is created equal - the design, conduct, analysis, and reporting of trials impact the validity of results. Argumentation quality assessment can help appraise the strength of claims and rationales in trials by examining qualities like cogency, effectiveness, and reasonableness. This assists reviewers in critically analyzing the literature rather than taking reported conclusions at face value. Furthermore, argumentation quality assessment techniques have great potential to assist clinical medical education. Medical residents must learn to construct high-quality clinical explanations regarding diagnoses, treatments, and recommendations. By automatically evaluating the reasoning quality of residents' explanations, the techniques explored in this thesis could give formative feedback to improve their argumentative writing abilities. This enables sharpening skills in evidence-based rationale construction that are essential for future medical practice. Argumentation quality assessment, therefore, has wide applicability spanning EBM, medical dialogue systems, and clinical education. However, adapting computational argumentation methods to the intricacies of clinical evidence requires specialized techniques tailored to the medical domain. The research presented in this thesis helps close this gap by pioneering quality assessment innovations needed for EBM applications.

### 2.3.2 External Medical Knowledge for Quality Assessment

Evaluating the quality of medical argumentation benefits greatly from incorporating domain knowledge from structured external sources. Medical knowledge bases provide statistical, relational, and definitional data that enable a richer assessment of clinical argumentation. Domain knowledge infusion allows moving beyond purely textual approaches to leverage insights from real-world clinical data when evaluating explanation quality.

A key resource is the Unified Medical Language System (UMLS) from the U.S. National Library of Medicine [49]. The UMLS integrates over 60 families of biomedical vocabularies, including clinical terminologies like SNOMED-CT and reference taxonomies like the NCBI organism taxonomy. It contains over 2.5 million biomedical concepts interconnected by 12 million relations. The UMLS provides a unified terminology framework to represent medical knowledge computationally.

While the UMLS offers broad terminology coverage, the Human Phenotype Ontology (HPO) [50] targets the specifics of human disease phenotypes. The HPO provides a standardized vocabulary of phenotypic abnormalities encountered in genetic disorders. Diseases in HPO are annotated with the symptoms, phenotypic features, and comorbidities associated with that condition. Critically, the linkages between diseases and phenotypes are supplemented with statistical data on the frequency of each symptom's occurrence according to aggregated patient data. For instance, a symptom may be labeled as very frequent (80-99% cases), frequent (30-79% cases), or

occasional (5-29% cases) for a particular disease. The richness of knowledge in HPO supports assessing the relevance of clinical findings or symptoms invoked when explaining a diagnosis. A highly frequent, obligate symptom provides stronger support for a disease hypothesis than an occasional, tangential one. By grounding analysis in the probabilistic data of HPO, computational methods can judge explanation quality from an evidence-based perspective. Medical dialogue systems could leverage this capability to offer data-driven feedback on clinician trainee diagnoses.

## 2.4 Natural Language Representations

Natural language processing (NLP) aims to enable machines to understand human language. While early NLP systems relied on hand-crafted symbolic rules, statistical and machine-learning approaches have become increasingly important [51–53]. Such approaches rely on training a mathematical model from the available data to be able to make predictions on new examples based on what was observed in the sample data. A core challenge in NLP is representing human language numerically so that machine learning models can process it [51, 52]. This quantification is essential for tasks like machine translation and text classification [52]. However, converting language into numbers is difficult. Languages have diverse vocabularies and grammar [54], so models designed for one may not work for others. Moreover, language is context-dependent and ambiguous. The meaning of a word or sentence depends heavily on the speaker’s intent and the discourse context [55]. Even humans sometimes misunderstand each other, highlighting the complexity of language understanding. Despite advances, fully understanding natural language remains an open challenge in NLP. Representing language numerically while preserving its nuance and context dependence is an active area of research. However, in the last decade, substantial progress has been made in natural language representation. The next section will showcase the major approaches to tackle this problem.

### 2.4.1 Context-free Representations

Context-free representations encode language without considering the surrounding context. Several major techniques are described below:

**Bag of Words** *The bag-of-words model (BOW)* represents text by the occurrences of words within a document. By retaining word counts, it captures the topics and themes present but discards word order and grammar. This approach was motivated by information retrieval tasks like document search, where keyword matching is important regardless of linguistic structure. The BOW model quantifies each sentence or document as a vector, where each dimension of the vector represents a word from the vocabulary. The value for each word in the vector representation is equivalent to the number of occurrences of that word in the sentence or document

being quantified. Consequently, this leads to high-dimensional sparse vectors, as most dimensions contain zeros and are meaningless for the numerical representation. Preprocessing techniques such as removing stopwords or lemmatization can help reduce the vocabulary size, but it remains enormous given the scale of modern corpora. Moreover, the lack of modelled word relationships means that semantic meaning is lost, synonyms cannot be identified and word sense disambiguation is not possible.

**TF-IDF** *Term frequency–inverse document frequency (TF-IDF)* builds on bag-of-words by weighting words based on their rarity, creating more meaningful vectors. Frequent words like “the” which appear across all documents are down-weighted, while rare words receive higher weights as they provide more specific information about document content. TF-IDF was developed to improve search engine results by emphasizing keywords that characterize relevant documents. However, like bag-of-words models, TF-IDF models lack semantic knowledge and cannot discern word meanings.

**N-grams** The local context, i.e., the surrounding words, plays an important role in understanding the semantics of text units. This contextual information is not captured by previous techniques like bag-of-words, which consider each word independently. To incorporate local context, the N-gram technique proposes encoding not just individual words but also co-occurring word combinations. N-grams were motivated by the need to handle multi-word phrases and idioms where meaning relies on word co-occurrence patterns. For example, properly interpreting “not happy” requires modeling bigrams. This approach can also help detect negations and disambiguate word senses based on context. The most common N-gram variants are bigrams (N=2) and trigrams (N=3), which model minimal local context. While most idioms and negations can be captured with low N, higher-order N-grams can be implemented to encode broader context. However, as with bag-of-words, the vector dimensionality grows exponentially with N, leading to sparsity issues. N-grams can also serve as a basic statistical language model, estimating the probability of a word given its context. The same preprocessing techniques used for bag-of-words, like stopwords and lemmatization, are applicable to N-grams. TF-IDF weighting can also be employed to emphasize informative N-grams. N-grams provide a simple method to incorporate local context, but they lack mechanisms to model long-range semantic dependencies. Long-range semantic dependencies refer to relationships between words that are not within a small local context window. For example, properly resolving the referent of a pronoun or resolving lexical ambiguity often requires reasoning about non-local context from earlier in the discourse. As N-grams only model a limited word window, they cannot capture these long-range dependencies that rely on broader discourse phenomena. Extensions like skip-grams have been

proposed to partially address this limitation, but fully modelling complex semantic relationships remains challenging with N-gram approaches.

**Word Embeddings** The mentioned techniques propose a way to encode documents or sentences in their entirety, but representations of individual words are also important. A basic approach is *one-hot encoding*, representing each word as a sparse vector with a 1 at the index for that word and 0s elsewhere. Representing words as unique one-hot encoded vectors leads to sparse high-dimensional representations unsuitable for neural networks. To address this, dense *word embeddings* were developed to represent each word as a lower-dimensional vector encoding semantic relationships based on distributional patterns. Word embedding models observe words in context across large corpora to learn vector representations, typically with a fixed size of 300 dimensions. These techniques create a vector space where words are positioned based on their semantic meaning, on which synonyms are clustered together, and relationships are modelled. For example, the vector representation of “King” is to “Man” in the same way that the vector representation of “Queen” is to “Woman”, capturing analogical reasoning. Any mathematical operations can be done with such vectors, allowing us to calculate “King” – “Man” to then add the vector of “Woman”, obtaining a vector result in which the closest word corresponds to “Queen”. Word embeddings operate well, even with polysemous words. The reason, as explained by Neelakantan et al. [56], is that “[i]n moderately high-dimensional spaces a vector can be relatively “close” to multiple regions at a time.”. However, biases in the training data affect embeddings. If we were to train our embeddings with a corpus based on news articles, the vector representation of the word “apple” may encode corporate rather than fruit meanings. Furthermore, only words appearing in the training data receive embeddings, causing out-of-vocabulary issues on unseen words.

Pre-trained word embeddings provide an advantage for machine learning approaches. Models like *Word2Vec* [57] or *GloVe* [58] are trained on diverse generic corpora to learn broadly useful representations that transfer across tasks. This is valuable when task-specific training data is limited. However, for specialized domains like biomedicine, pre-trained embeddings may perform poorly if the vocabulary and semantics differ greatly from the general domain. In such cases, custom in-domain embeddings can be learned from texts in that domain, capturing nuanced meanings and terminology. This domain-specific fine-tuning typically improves performance but requires sufficient task data. Pre-trained embeddings provide a useful starting point when data is scarce. With more data, fine-tuning in-domain corpora yields embeddings better tailored to the task. Various types of word embeddings were experimented with in this thesis as input representations for neural networks and other machine learning models.

## 2.4.2 Contextualized Representations

Recent state-of-the-art textual representations utilize *contextualized embeddings*, which address limitations of static word embeddings. While static embeddings capture general semantic information, they ignore surrounding context and learn only one fixed vector per word. This representation does not adapt based on context, leading to issues with polysemy. Words with multiple meanings are conflated into a single vector, requiring downstream models to disambiguate meanings. However, properly resolving word senses necessitates modeling context. For example, in sentiment analysis, correctly classifying the polarity of “The bank closed my account” requires understanding whether “bank” refers to a financial institution or a river bank based on context. Static embeddings cannot effectively disambiguate these two meanings, demonstrating a key limitation. The context surrounding a word determines its intended semantics, which should be encoded in its representation. In contrast, contextualized embeddings are “dynamic”, producing context-aware representations tailored to each instance. In the case of our example, contextualized embeddings can encode the word “bank” differently based on the surrounding text [59, 60], providing the disambiguation needed for accurate sentiment analysis. Rather than learning one vector per word, they encode words differently based on the linguistic context in each case. This better reflects how word usage and meaning varies across different contexts.

**ELMo** Peters et al. pioneered contextualized word representations through Embeddings from Language Models (ELMo) [61]. Unlike static embeddings, ELMo creates dynamic word representations based on the entire input sentence. Under the hood, ELMo employs a bidirectional language model (biLM) architecture comprising two separate but linked LSTM-based neural networks [62]. The first is a forward LSTM that processes the sentence left-to-right, while the second is a backward LSTM that processes right-to-left. Each LSTM layer captures contextual information from one direction, learning long-range dependencies within the sequence. These LSTMs are pre-trained on a large text corpus to perform language modeling - predicting the next word in context. The internal representations from the forward and backward LSTMs are then concatenated depthwise to produce the ELMo word vectors. Concatenating bidirectional outputs allows ELMo to incorporate both past and future context when constructing each word embedding. Moreover, combining representations from multiple LSTM layers captures both low-level syntactic and higher-level semantic information. Consequently, ELMo generates rich dynamic word vectors tailored to each context, significantly advancing over previous static embeddings.

**Transformer** The Transformer model, proposed by Vaswani et al. [63], significantly shifts the paradigm in designing architectures for NLP tasks. Unlike its predecessors, the Transformer model introduces the concept of self-attention, or scaled dot-product attention, which allows the model to consider the relevance of each word



in a sentence for encoding a particular word. This capability enables the model to effectively capture the context of the sentence. In contrast to recurrent models such as LSTMs that process sentences sequentially, Transformers process all words in the sentence simultaneously, paving the way for more parallel computation and handling long-range dependencies more effectively. The Transformer model comprises two main components: an encoder that processes the input text and a decoder that generates the output text word by word. Each component consists of multiple layers of self-attention mechanisms and feed-forward neural networks, allowing the model to learn complex patterns in the data. Notably, the decoder generates each word considering the encoder's output and its previously generated words. This architecture has become foundational for many subsequent models, including BERT and GPT. The architecture is illustrated in Figure 2.3

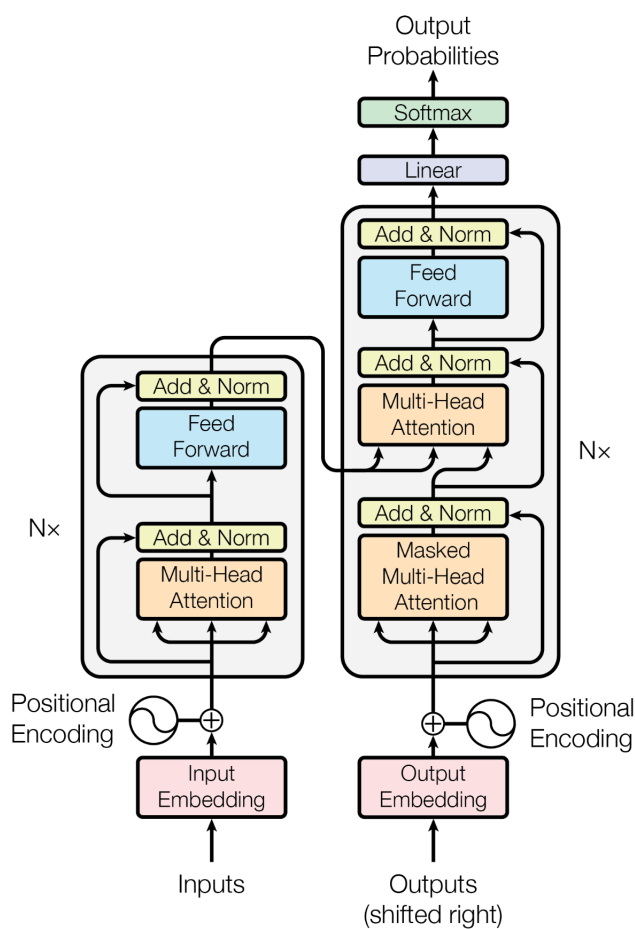


FIGURE 2.3: The transformer model architecture. Figure drawn from [63].

**GPT** Radford et al. proposed the Generative Pretrained Transformer (GPT) [64], another influential contextual embedding technique. Utilizing the decoder of the Transformer model, GPT is pre-trained with a unidirectional language modelling objective. This allows it to effectively generate text, proving particularly effective in tasks that involve text generation. However, GPT's unidirectional nature means it

can consider context only from one direction when generating embeddings. While this is sufficient for many applications, it can limit the model’s ability to fully understand the semantic nuances of language. This limitation is addressed by the subsequent development of the BERT model.

**BERT** Devlin et al. [21] introduced the Bidirectional Encoder Representations from Transformers (BERT) architecture, a significant advancement in the realm of contextualized embeddings. Distinct from earlier models like ELMo or unidirectional models like GPT, BERT utilizes the Transformer encoder in a bi-directional manner. This characteristic allows BERT to understand the context of a word from both directions, i.e., from the words preceding and following it, making it particularly effective in disambiguating word meanings based on their context. BERT is pre-trained using two main objectives: masked language modelling (MLM) and next sentence prediction (NSP). The MLM objective is inspired by the Cloze task [65], where random words in the input are masked, and the model is tasked with predicting the original masked words. Unlike the unidirectional next-word prediction task, the MLM objective allows BERT to take into account both the left and the right context. The second pretraining task, NSP, tests the model’s ability to understand the relationship between two sentences. This is particularly useful for many downstream NLP tasks that require an understanding of sentence relationships, such as Question Answering. In the NSP task, the model is given two sentences and must predict whether the second sentence follows the first in the original text. After pretraining, BERT can be *fine-tuned* on specific downstream tasks. Fine-tuning involves continuing the training process on a specific task (like sentiment analysis or question answering) using a smaller, task-specific dataset. This process of fine-tuning adjusts the pre-learned representations to better suit the specific task, leveraging the broad language understanding learned during pretraining while adapting to the task-specific patterns. Fine-tuning is relatively inexpensive compared to pretraining, making BERT a versatile and efficient model for a variety of NLP tasks

The versatility and effectiveness of the BERT model have led to the development of numerous specialized models that build upon the original architecture or adapt it to specific domains. These models leverage the power of BERT while tailoring its capabilities to better suit specific tasks or types of data. For instance, domain-specific models such as SciBERT [66], BioBERT [67], and PubMedBERT [68] are trained on scientific, biological, and medical texts, respectively. These models are able to capture domain-specific language and semantics, greatly improving performance on tasks within these fields. In addition to these domain-specific models, there have also been architectural advancements that improve upon the base BERT model. For example, RoBERTa [69] modifies the BERT training procedure to improve its performance, including training with larger mini-batches, using more data, and removing the next sentence prediction task. ALBERT [70], on the other hand, reduces the

model size of BERT while maintaining comparable performance, making it more efficient. These models demonstrate the ongoing evolution of BERT and its continued influence in the field of natural language processing.

**SBERT** Expanding the scope of contextualized models beyond word-level embeddings, Sentence-BERT (SBERT) [71] revolutionizes the approach towards understanding sentence-level semantics. While models like BERT generate embeddings for individual words, SBERT takes a step further to create embeddings for entire sentences. This is accomplished by modifying the BERT architecture to accept pairs of sentences and training it on Natural Language Inference (NLI) tasks. In these tasks, the model learns to classify pairs of sentences into categories such as 'entailment', 'contradiction', and 'neutral', effectively teaching the model about semantic relationships between sentences. The SBERT model thus learns to encode the context and semantic relationships between the words within each sentence, as well as the relationships between different sentences. The resulting sentence embeddings encapsulate the overall semantic content of each sentence, enabling efficient semantic similarity comparisons between sentences.

This approach represents a shift in focus from understanding individual words to understanding the larger units of meaning in language, providing a more holistic view of textual data. Applications that require understanding the overall meaning of sentences, such as semantic search, text clustering, and paraphrase detection, can greatly benefit from SBERT embeddings.

## Chapter 3

# Argumentation Quality

*This chapter introduces my first contribution towards the assessment of argumentation quality. More precisely, my contribution is twofold: first, I present a novel resource of 402 student persuasive essays, where three main quality dimensions (i.e., cogency, rhetoric, and reasonableness) have been annotated, leading to 1908 arguments tagged with quality facets; second, I address the challenging task of argumentation quality assessment proposing a novel neural architecture based on graph embeddings, combining both the textual features of the natural language arguments and the overall argument graph, i.e., considering also the support and attack relations holding among the arguments. Finally, a discussion of the obtained results and limitations of this approach is presented. This chapter comprises the work published at the International Conference on Empirical Methods in Natural Language Processing (EMNLP-2022) [20] and at the European Conference on Argumentation (ECA-2022) [72].*

Argumentation is the process by which arguments are constructed, compared, evaluated in several respects and judged in order to establish whether any of them is warranted. Argumentation is an effective approach for solving various theoretical and practical problems [1, 73], like explaining and justifying the decision-making outcomes and reasoning under inconsistent and incomplete information.

A major component of the argumentation process concerns the assessment of a set of arguments and of their conclusions to establish their justification status, and therefore compute their acceptability degree [36]. Both qualitative and quantitative approaches have been proposed in the literature to assess the acceptance of an argument. However, the assessment of the argument's acceptability is only a (basic) part of the complex assessment tasks required in argumentative processes in many everyday life applications and contexts, e.g., in medicine and education.

The issue of assessing an argumentation is particularly critical when considering the different aspects of artificial argumentation, from the identification of real natural language arguments and their relations in text, to the computation of the justification status of abstract arguments [36], to the gradual assessment of arguments [74, 75] based, e.g., on the trustworthiness of the argument proponents [76] or

on the value promoted by the argument [77]. In particular, despite some approaches addressing the automatic assessment of natural language arguments [4, 5, 78], this issue remains largely unexplored and unsolved.

In this chapter, I address this open issue and answer the following research questions: (i) How can we model the multi-dimensional notions of quality to capture key aspects like cogency, rhetoric, and reasonableness? and (ii) Can integrating textual features with graph embeddings and emotion detection improve the assessment of argumentation quality dimensions like cogency, rhetoric and reasonableness?

More specifically, I propose an argument mining [26, 79, 80] approach to identify and classify natural language arguments along with quality dimensions. First, I define and annotate three prominent quality dimensions for natural language argumentation, i.e., *cogency*, *rhetoric* and *reasonableness*, on an existing dataset of student persuasive essays [81]. Methods for assessing each quality dimension are then proposed, including SVMs and Random Forest with various word embeddings and fine-tuned transformer models empowered with graph embeddings and emotion detection to address the task.

The work I present in this chapter is motivated by the lack of existing resources of natural language argumentation annotated with quality dimensions and the need for effective methods to address this task. This contribution advances the state of the art with a novel resource and an effective method.

This chapter is organised as follows: in Section 3.1 I introduce existing work on the quality evaluation of natural language arguments. In Section 3.2 I present the resource of natural language argument graphs annotated with the quality dimensions, while in Section 3.3 I describe the proposed architecture to automatically assess these quality dimensions on the bipolar argumentation graphs. In Section 3.4 I discuss obtained results, and conclusions at the end of the chapter.

### 3.1 Related Work

Recent approaches in Argument(ation) Mining (AM) [26, 79, 80] tackle specific argument qualities features, such as argument relevancy [82], convincing arguments [41] and overall argument quality [42]. Previous work on student essays aimed to assess clarity [83], organization [44] and argument strength [84]. [85] target the automatic prediction of the quality of student reflective responses, showing how expert-coded quality ratings and quality predictions based on their features positively correlate with student learning gain.

Defining the characteristics of a good and successful argument is a hard task. Different approaches have been proposed to assess the logical, rhetorical, and dialectical quality dimensions of natural language arguments.

Wachsmuth et al. [4] derive a taxonomy of argumentation quality that systematically decomposes quality assessment based on the interactions of 15 widely accepted

quality dimensions. The three main characteristics are *Cogency*, *Effectiveness* and *Reasonableness*. As a follow-up, [78] investigate how effectively each dimension can be automatically assessed, modelling features such as content, style, length and subjectivity. This text-only assessment yields moderate learning success for most of the evaluated dimensions. In another text-only approach, Lauscher et al. [86] describe a large argument quality corpus with data extracted from forums. They propose the first computational model to automatically evaluate Cogency, Reasonableness, Effectiveness and overall quality.

Saveleva et al. [5] presents an argument quality assessment method defined as a graph classification task. The authors reconstruct the graph structure of the arguments within the argument quality dataset presented by Wachsmuth et al. [4], showing that this is feasible only in some cases. The reconstructed structures are composed of claims and evidence connected by a support relation, disregarding important elements like counterarguments and rebuttals. Results indicate that discourse-based argument structures reflect the qualitative properties of the arguments. For rhetorical aspects, Duthie et al. [87] show the impact of the different rhetorical strategies used in political discourse. For Automatic Essay Scoring, Zhang et al [88] show how human-labelled evidence scores can be replaced with other automated essay quality signals, such as word count and topic distribution similarity.

In this thesis, we advance the state of the art of natural language argument quality assessment by investigating three main quality properties of *persuasive essays* grounding on social science argument quality assessment scores [19]. Moreover, we propose a novel method to evaluate the reasonableness of an argument by combining cogency properties with the argumentation graph structure.

## 3.2 Quality dimensions of persuasive essays

To annotate the quality dimensions of persuasive essays, we rely on the corpus built by Stab and Gurevych [81], containing 402 persuasive essays annotated with the argument components (i.e., evidence, claims and major claims) and relations (i.e., support or attack). This results in 402 argument graphs where the argument components are the nodes of the graph, and the argumentative relations are the edges of the graph. We add a new annotation layer by manually labelling for each argument in the essays the following three quality attributes: *cogency*, *reasonableness* and *argumentation rhetoric*, following the taxonomy proposed by [4]. Taking advantage of the relation annotations, we use the *argument graph* (i.e., argument components and their relations) to assist annotators in their annotation process.

### Annotation guidelines.

Defining the characteristics of a good and successful argument is a hard task. First, we must address the several text rating procedures that have been proposed in the literature. Different factors, such as the aim of the assessment, the freedom given

to the raters, and the number of texts to be analyzed should be considered when evaluating the quality of argumentative texts. Following Coertjens et al. [89], rating procedures can be classified into two dimensions: Holistic vs. analytic and absolute vs. comparative. Holistic rating entails evaluating texts as a complete entity, while analytic rating involves assessing multiple text features. In absolute ratings, each text is assessed based on predefined criteria or description, while in comparative ratings, texts are compared to each other to determine their score. In this study, our objective is to assess the quality of argumentative texts in persuasive essays through the application of a consistent and absolute analytic rating system. We focus on assessing the quality of persuasive essays, where the aim is to assess each essay on its own. The aim is to ensure that the evaluation is based solely on the essay's content, rather than the subjective bias of the evaluator and that the assessment results are consistent across all raters.

A frequently used analytic rating procedure is the use of rubrics. A set of criteria for grading text is defined within the rubrics and they usually contain evaluative criteria, quality definitions for those criteria at particular levels of achievement, and a scoring strategy. In an analytic rubric, the text features are pre-determined. However, the significance or weight assigned to each feature is not always determined in advance. As established by Sasaki et al. [90] and Coertjens et al. [89], evaluators are given the flexibility to independently assign weights to the pre-determined text features. Consequently, variations in assessments of a single text among evaluators may arise.

To tackle this issue, Stapleton and Wu [19], describe the weight of the separate text features in a rubric as fixed. In this rubric, the authors stated that a strong argumentative text is composed of two important elements. (i) an argumentative text must be constructed taking into account all elements contributing to a good *quality of argumentation* and (ii) attention must be paid to the *quality of the content* of the text. This rubric contemplates several characteristics of the standard definition of Cogency and Reasonableness, such as Relevancy, Acceptability, and Soundness as well as the presence of counterarguments and rebuttals. Given that our goal is to assess persuasive essays written by students, we rely on the quality evaluation process proposed by Stapleton and Wu [19]. Tables 3.1, 3.2 and 3.3 show the analytic scoring rubrics proposed by Stapleton and Wu [19]. A scale of 0, 10, 15, 20, and 25 is given to assess the Cogency and Reasonableness of a given argument.

**Cogency.** An argument should be seen as cogent if it has individually acceptable premises that are relevant to the argument's conclusion and that are sufficient to draw the conclusion [4]. Annotators were provided with Table 3.1 to assess the cogency dimension. Following this definition, we define the *acceptable* premises as the ones that are worthy of being believed, and the *relevant* ones as those that contribute to the acceptance or rejection of the argument's conclusion. These criteria are considered in point (b) (Table 3.1) whilst the structural information about the argument graph is addressed in point (a). Example 3.2.1 shows the cogency annotation on a

persuasive essay extracted from the dataset proposed by Stab and Gurevych [81]. The first sentence is the major claim, while the claim to be assessed is in bold and the premises supporting it are in italics. Example 3.2.1 is annotated with a cogency score of 25, given that the author presents multiple premises which are acceptable and relevant to draw a conclusion. On the other hand, Example 3.2.2 is annotated with a cogency score of 0 (lower score) given that the author provides no acceptable premises for her claim.

Score: 25	Score: 20	Score: 15
a. Provides multiple reasons for the claim(s), and b. All reasons are sound/acceptable and free of irrelevancies	a. Provides multiple reasons for the claim(s), and b. Most reasons are sound/acceptable and free of irrelevancies, but one or two are weak	a. Provides one to two reasons for the claim(s), and b. Some reasons are sound/acceptable, but some are weak or irrelevant
	Score: 10	Score: 0
	a. Provides only one reason for the claim(s), or b. The reason provided is weak or irrelevant	a. No reasons are provided for the claim(s); or b. None of the reasons are relevant to/support the claim(s)

TABLE 3.1: Analytic Scoring Rubric to assess Cogency [19].

**Example 3.2.1** *We should attach more importance to cooperation during primary education. [Through cooperation, children can learn about interpersonal skills which are significant in the future life of all students] <sub>1</sub>. [What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others] <sub>1</sub>. [During the process of cooperation, children can learn about how to listen to opinions of others, how to communicate with others, how to think comprehensively, and even how to compromise with other team members when conflicts occurred] <sub>2</sub>. [All of these skills help them to get on well with other people and will benefit them for the whole life] <sub>3</sub>.*

**Example 3.2.2** *It's certainly better for children to grow up in a big city. [Growing up in the countryside is not such a good experience] <sub>1</sub>, [you won't know a lot of people, there are gossips everywhere, and your life will be really limited]. <sub>1</sub>*

**Reasonableness.** An argumentation should be seen as reasonable if it contributes to the resolution of the given issue in a sufficient way that is acceptable to the target audience [4]. The Analytic Scoring Rubric for Reasonableness (Tables 3.2 and 3.3 [19]) integrates these concepts and follows the idea of evaluating the argumentation graph with a focus on the counterarguments and their respective rebuttals. Annotators were asked to annotate both the Reasonableness Counterargument and the Reasonableness Rebuttal whenever an essay presented a counterargument. Whilst the definitions of Reasonableness and Cogency are similar, the key difference is that with Cogency we evaluate the premises of the argument and with Reasonableness



the whole argumentation graph involving the argument to be assessed (including its counterarguments and their rebuttals).

Assessing the Reasonableness of an argument implicates analysing its counterarguments and the related rebuttals. In the example of Figure 3.1, we assess the reasonableness quality dimension following Tables 3.2 and 3.3: for counterargument *Claim E*, we can see that no reasons, or premises, provided to support it. This falls under the criteria for Score 0 for Reasonableness Counterargument. For the rebuttal, we can see that *Claim F* correctly points out the weakness of the counterargument, providing an acceptable and sound premise and with a reasoning quality stronger than of the counterargument, therefore falling under the criteria for Score 25.

Score: 25	Score: 20	Score: 15
a. Provides multiple reasons for the counterargument claim(s) /alternative view(s), and b. All counterarguments/reasons for the alternative view(s) are sound/acceptable and free of irrelevancies	a. Provides multiple reasons for the counterargument claim(s)/alternative view(s), and b. Most counterarguments/reasons for the alternative view(s) are sound/acceptable and free of irrelevancies, but one or two are weak	a. Provides one to two reasons for the counterargument claim(s) /alternative view(s), and b. Some counterarguments/reasons for the alternative view(s) are sound/acceptable, but some are weak or irrelevant

Score: 10	Score: 0
a. Provides only one reason for the counterargument claim(s)/alternative view(s), or b. The counterargument/reason for the alternative view is weak or irrelevant	a. No reasons are provided for the counterargument claim(s)/alternative view(s); or b. None of the reasons are relevant to/support the counterargument claim(s)/alternative view(s)

TABLE 3.2: Analytic Scoring Rubric for assessing Reasonableness Counterargument [19].

**Argumentation Rhetoric.** Annotators were asked to evaluate at the argument level which rhetoric strategy the argument is following among *ethos*, *logos*, and *pathos* [3]. *Logos* is the act of appealing to the audience through reasoning or logic, by citing facts and statistics, historical and literal analogies. *Ethos* is the act of appealing to the audience through the credibility of the author’s beliefs or authority. *Ethos* can be applied by choosing the appropriate language for the audience and the topic (e.g., the proper level of vocabulary), making the author sound fair or unbiased, introducing her expertise, and using correct grammar and syntax. *Pathos* means to persuade an audience by appealing to their emotions. Some examples of persuasive essays annotated with argumentation rhetoric are available in the Appendix. Authors use *pathos* to invoke sympathy from an audience, and to make the audience feel what the author wants them to feel. A common use of *pathos* would be to draw pity from

Score: 25	Score: 20	Score: 15
a. Refutes/points out the weaknesses of all the counterarguments, and b. All rebuttals are sound/acceptable c. The reasoning quality of all the rebuttals are stronger than that of the counterarguments	a. Refutes/points out the weaknesses of all the counterarguments, and b. Most rebuttals are sound/acceptable, but one or two are weak c. The reasoning quality of most rebuttals are stronger than that of the counterarguments, while one or two are equal to that of the counterarguments	a. Refutes/points out the weaknesses of all the counterarguments, and b. Some rebuttals are sound/acceptable, but some are weak c. The reasoning quality of some rebuttals are stronger than that of the counterarguments, while some are weaker than that of the counterarguments
	Score: 10	Score: 0
	a. Refutes/points out the weaknesses of some counterarguments, or b. Few of the rebuttals are sound/acceptable; most of them are weak, or c. The reasoning quality of most rebuttals are weaker than that of the counterarguments	a. No rebuttals are provided; or b. None of the rebuttals can refute the counterarguments

TABLE 3.3: Analytic Scoring Rubric for assessing Reasonableness Rebuttal [19].

the audience, and can be induced by using meaningful language, emotional tone, emotion-evoking examples and implied meanings.

We now describe Argumentation rhetoric dimension through some examples from the persuasive essays dataset [81] with the help of Examples 3.2.3, 3.2.4 and 3.2.5.

In Example 3.2.3 the claim (in bold) appeals to emotions *Pathos* when the author describes how “people are better taken care’ in premises 1 and 3 (in italic). In Example 3.2.4 the authors employ *Ethos*, we can notice that the author refers to personal experiences in premises 1 and 2. Example 3.2.5 employs *Logos*, the author refers to a formal study, in premise 2, in order to support its claim.

**Example 3.2.3** *The advanced medical care brings with it more benefits than disadvantages. [The main advantage of high tech medical care is that people are better taken care so that they have a good health]<sub>1</sub>. [Healthy workers can create more productivity]<sub>1</sub> [They can contribute effectively to the development of the economy]<sub>2</sub>. [They do not have to spend more time in health checking or treatment]. <sub>3</sub> [this saves an amount of time as well as cost]<sub>4</sub>.*

**Example 3.2.4** *People should sometimes do things that they do not enjoy. [In personal live, we have some responsibilities towards to other people, there is nobody who likes all of these responsibilities]<sub>1</sub>. [Housework is very difficult for me, although my husband helps me some of them, but it is my responsibility]<sub>1</sub>. [I really don’t like any of them,*

however I should do]<sub>2</sub>, [most people's lives are filled with tasks that they don't enjoy doing]<sub>3</sub>.

**Example 3.2.5** *Following celebrities can be dangerous for the youth. [This has an overall effect on personality and future of an individual, following celebrities blindly affects the health of adolescents.]<sub>1</sub> [Many young people indulge themselves in drugs and start smoking at an early age]<sub>1</sub>. [In a survey carried out in a university, it was asked to students that why did they start smoking, then around forty percent of individuals answered that they wanted to look like their favorite screen actor while smoking cigarettes]<sub>2</sub> [Imitating celebrities has a negative influence on health of young individuals]<sub>3</sub>.*

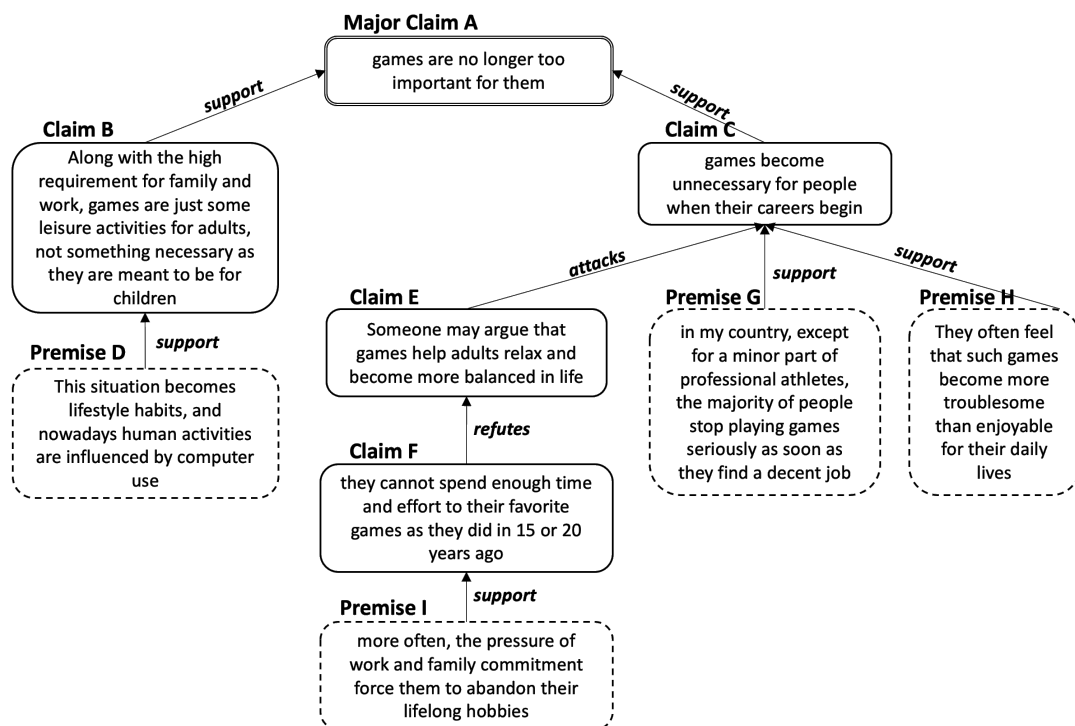


FIGURE 3.1: Example of an argument graph of a persuasive essay [81].

Before starting the annotation process, three annotators (English speakers and experts in Argumentation Mining) carried out a training phase, during which they studied the guidelines and discussed about the ambiguities between the scores for the definitions of Cogency and Reasonableness, amongst others. Then, the annotators were presented with an argument from a persuasive essay (a Claim or Major Claim component) and its full argument graph, and they had to annotate the argument quality following the rubric scores. To prove the reliability of the annotation task, the inter-annotator agreement (IAA) has been calculated on an unseen set of 33 essays, obtaining a Fleiss' kappa of 0.68 for Cogency, 0.78 for Reasonableness Counterargument, 0.84 for Reasonableness Rebuttal and 0.85 for Argumentation Rhetoric.

Despite this substantial agreement, an issue for the annotators was the difficulty to opt for a precise score, like 25 or 20. To minimize subjectivity issues in the manual annotation and the consequent noise in the training and testing phases for the automatic assessment of these scores, we decided to merge Score 25 with Score 20, and Score 15 with Score 10, reducing the number of labels to 3 (Score 0 is kept as is). We then proceeded to recompute Fleiss’ kappa score, obtaining an increment for Cogency (from 0.68 to 0.86) only. For this reason, we decided to rely on a three-label score for Cogency prediction (i.e., 0, 15, 25), and to keep the more fine-grained score for Reasonableness (i.e., 0, 10, 15, 20, 25). The annotators performed then a reconciliation phase, during which they discussed to reach an agreement on the cases of disagreements. The rest of the annotation was carried out by one of the expert annotators. Tables 3.4 and 3.5 report on the statistics of the final dataset<sup>1</sup>.

Score	Cogency	Reas. Counterargument	Reas. Rebuttal
0	19.70%	27.27%	79.82%
10	9.38%	25.45%	9.65%
15	19.14%	26.36%	4.39%
20	31.71%	13.64%	3.51%
25	20.08%	7.27%	2.63%

TABLE 3.4: Statistics of the dataset, reporting on the percentage of Cogency and Reasonableness for each score.

No Rhetoric	Ethos	Logos	Pathos
76.04%	11.51%	6.79%	5.66%

TABLE 3.5: Statistics of the dataset, reporting on the percentage and type of Rhetorical arguments.

### 3.3 Automatic assessment of argumentation

An overview of the automatic argument quality assessment framework we propose is visualized in Figure 3.2. Starting from the persuasive essays where argument components and their relations are identified, the goal is to assess the quality of each argument (i.e., the quality of each claim). Three scores are computed: a *cogency* score in the range  $\{0, 15, 25\}$ , an *argumentation rhetoric* label among *ethos*, *logos*, and *pathos*, and a *reasonableness* score in the range  $\{0, 10, 15, 20, 25\}$ . Two different methods are combined to effectively assess the quality dimensions of the arguments: (i) the cogency score and the argumentation rhetoric labels are predicted using an attention-based neural architecture which employs the argumentation graphs through graph embeddings, and (ii) the reasonableness score is computed by means of an algorithm, combining the cogency score predicted at step (i) and the graph structure of

<sup>1</sup><https://gitlab.com/santimarro/persuasive-essays-argument-quality-dataset>

each persuasive essay. In the following, we present the features we extracted from the persuasive essays to predict the cogency score and argumentation rhetoric labels, the neural architecture we define to predict these two quality dimensions, and we conclude with the reasonableness algorithm used to assess this score.

### 3.3.1 Cogency and rhetoric scoring assessment

Different techniques [4, 78] have been proposed to automatically assess the cogency of arguments through text-based methods alone. However, following the definition of cogency by Blair [2], a cogent argument requires premises that are not only acceptable, relevant and sufficient based on their textual content, but also structured in a way that properly supports the conclusion. Specifically, the presence of multiple premises that collectively demonstrate acceptability, relevance and sufficiency is a key structural consideration for cogency.

This need for assessing cogency using both textual and structural features is reinforced by the rubric of Stapleton and Wu [19], where the cogency score is determined by a combination of (a) the number of premises supporting a claim, and (b) the acceptability, sufficiency and relevancy of each premise based on the text.

To represent and leverage structural features like the number and organization of premises, we propose using graph embeddings. Graph embeddings are vector representations that encode the topology and connections in a graph network. For argument graphs, where nodes are claims and premises and edges indicate support/attack relations, graph embeddings can capture informative structural properties. This includes the number of premises attached to a claim, connectivity patterns between premises, and relative position in the graph. Such structural nuances can complement the text-based acceptability, sufficiency and relevancy assessment. Simple text methods that look at premises individually would miss these graph-level insights. By combining text embeddings of argument components with graph embeddings of the full structure, we can account for both textual and structural influences when predicting cogency scores. This allows modeling cogency as the rubric defines it - based on collective premises and their relations, not just isolated texts. Our approach leverages recent graph embedding techniques like FEATHER-G [91] to compute argument graph representations.

For textual feature generation, we employ a variety of embedding techniques from static word vectors like GloVe [58] to contextualized models such as BERT [21] and Longformer [92]. To represent each argument component, we concatenate all the sentences composing the claim itself along with any related premises or claims linked via the argument graph structure. Since combining these sentences results in lengthy documents, we utilize Longformer given its ability to process sequences up to 4096 tokens while maintaining state-of-the-art performance. For graph feature generation, we adopt the FEATHER-G framework [91], which combines node attribute information and random walk weights to describe node neighborhoods.

These node-level representations are pooled using mean pooling to obtain graph-level embeddings. FEATHER-G allows encoding both topology and node content into a single vector that represents the overall graph structure. By extracting features for both text and graphs, we can develop models that account for linguistic and structural aspects of argumentation. The contextualized text embeddings provide nuanced representations of component semantics and syntax. The graph embeddings complement this by capturing relational patterns and properties. Together, they allow us to predict cogency based on Blair’s multidimensional definition encompassing both textual acceptability and premise sufficiency.

Rhetorical strategies like ethos, logos and pathos rely on different means of persuasion. Specifically, pathos invokes strong emotions to persuade the audience. As such, automatically detecting pathos requires identifying the emotional content within arguments, beyond just the text. Recent sentiment analysis and emotion detection techniques [93–95] demonstrate progress in modeling the affective dimension of text. Building on these advances, we propose integrating emotion features into argument representations to better discern pathos from more neutrally-toned strategies like logos and ethos.

Concretely, we build emotion embeddings for the arguments by extracting emotion labels leveraging a transformer-based pre-trained model T5 [96], fine-tuned on the emotion recognition dataset by [97] for the Emotion Recognition downstream task. This approach allows us to obtain an emotion label amongst sadness, joy, love, anger, fear, or surprise. We then obtain a word embedding as a feature vector by either directly extracting the label representation from the fine-tuned model or employing the label to obtain a word embedding using GloVe. We hypothesize that arguments exhibiting high emotion intensity are more likely to represent pathos rhetorical attempts. To test this, we combine the extracted emotion features with text embeddings of arguments into a single representation. This allows rhetorical strategy classification to jointly consider linguistic style/content along with affective properties. The integrated text and emotion features help differentiate impassioned pathos arguments from logically-driven logos or credibility-focused ethos. By incorporating emotion detection into argument analysis, we provide a novel approach to rhetorical strategy identification tailored to the nuances of pathos. The ability to model persuasive appeals based on both text and emotionality of arguments is a unique advantage of our method.

After feature generation, we automatically assess each quality attribute. For Cogency and Reasonableness, Support Vector Machines (SVM) [98], Random Forests [99], Bidirectional LSTM-CRF [100] and fine-tuned Longformer [92] with an added dense layer for classification models were utilized. In our experiments, we evaluate different combinations of these methods with different combinations of the previously mentioned embeddings as an input vector.

As the majority of the arguments in our dataset have a non-rhetorical structure (Table 3.5), the automatic Argumentation Rhetoric assessment task was divided into

two different steps. First, a binary classification task to distinguish between a *rhetorical* and a *non-rhetorical* argument, and then a multi-label classification task to classify a rhetorical argument into *ethos*, *logos* or *pathos*. For both tasks, the implemented architectures are the same.

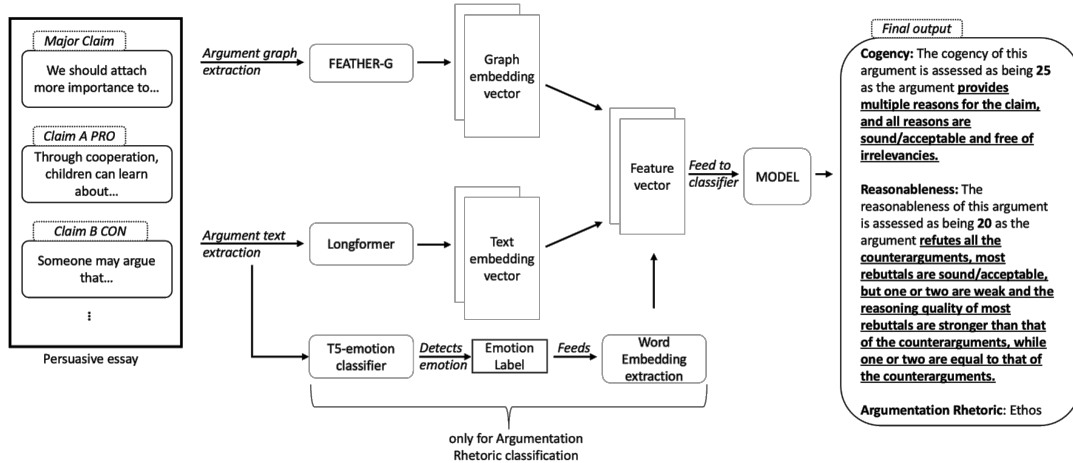


FIGURE 3.2: Overview of our natural language argumentation quality prediction model.

### 3.3.2 Reasonableness scoring assessment

Given the fact that in our dataset the majority of the essays did not present any counterarguments or, for a given counterargument, there was no rebuttal, our models did not have enough data to learn how to classify the reasonableness quality dimension. Motivated by this and by the consideration that the structure of the argumentation graph plays a main role in assessing reasonableness, we propose a novel approach to address this task. The reasonableness dimension [19] takes into account (i) the cogency of the counterarguments attacking the argument we want to assess the reasonableness of, (ii) the cogency of the rebuttals to these counterarguments (i.e., the arguments attacking the counterargument), and (iii) the relative number of rebuttals and counterarguments. This means that to effectively compute the reasonableness dimension, we need to combine the cogency-based quality of the argument components and the structure of the argumentation graph. We define the cogency function CV which assigns to each argument component A a cogency value in  $\{0, 10, 15, 20, 25\}$ , using the SVM plus graph embeddings approach we proposed.

Based on the rubric by Stapleton and Wu [19], we propose an algorithm (Algorithm 1) to compute the reasonableness score of the arguments in our argumentation graphs. In this Rebuttal Reasonableness Score algorithm, the reasonableness score of the argument component A is 0 if (i) no attack to the counterarguments in CA of A holds (line 19), or (ii) the cogency value of the argument components defending A, i.e., attacking the counterarguments of A, is 0 (line 16).

For the remaining reasonableness scores, the reasonableness score of A is 10 if (i) at least one and less than half of its counterarguments are attacked (line 25), or (ii)

the cogency score of more than half of the argument components defending it is 10 (line 22), or (iii) the cogency score of more than a half of the argument components defending it is lower than the cogency score of the counterarguments of  $A$  (line 29). The reasonableness of  $A$  is 15 if (i) all the counterarguments of  $A$  are attacked (line 32), and (ii) the cogency score of at least one of the argument components defending  $A$  is equal to or higher than 15 and at least one of the argument components defending  $A$  has a cogency score lower than 15 (line 33), and (iii) the cogency score of at least one of the argument components defending  $A$  is higher than the cogency score of the counterarguments of  $A$  and at least one of the argument components defending  $A$  has a cogency score lower than the cogency score of the counterarguments of  $A$  (lines 34 and 35, respectively). The reasonableness score of  $A$  is 20 if (i) all the counterarguments of  $A$  are attacked (line 32), and (ii) the cogency score of more than half of the argument components defending  $A$  is equal to or higher than 20, and at least one of the argument components defending  $A$  has cogency score equal to or lower than 10 (line 40), and (iii) the cogency score of more than half of the argument components defending  $A$  is higher than the cogency score of the counterarguments of  $A$  while one or two of the argument components defending  $A$  has cogency score equal to that of the counterarguments of  $A$  (line 41). Finally, the reasonableness score of  $A$  is 25 if (i) all the counterarguments of  $A$  are attacked (line 32), and (ii) the cogency score of all of the argument components defending  $A$  is 25 (line 45), and (iii) the cogency score of all of the argument components defending  $A$  is higher than the one of all of the counterarguments of  $A$  (line 46).

Let us consider the example in Figure 3.1. We aim to assess the reasonableness score of claim  $C$ . It holds that  $CV(E) = 0$  ( $E$  is the counterargument of  $C$ ) and  $CV(F) = 10$  ( $F$  is the rebuttal of  $E$ ). Starting from the cogency scores of all the counterarguments and rebuttals of our target argument component  $C$ , we can see that if the cogency value of every rebuttal is 10 (the cogency score of claim  $F$ ), then the reasonableness of claim  $C$  is 10.

After the automatic assessment of the Cogency, Rhetoric, and Reasonableness dimensions, the obtained scores are used to help the student to improve the essay. Our pipeline ends with the automatic generation of the scores using this template: The [QUALITY DIMENSION] of this argument is assessed as being [PREDICTED SCORE] as the argument [DEFINITION] (see Figure 3.2).

### 3.4 Evaluation

In the following, we report on the experimental setup, the obtained results and the error analysis.

**Experimental Setup.** For argument quality prediction, the embeddings (see Section 3.3) were combined with either (i) a Random Forest, (ii) a LSTM, (iii) a dense



---

**Algorithm 1** Rebuttal Reasonableness score
 

---

**Require:** Argument Component  $A$ ,  $CA$  a set with all the argument components that directly attack  $A$ .  
**Ensure:** Returns 0, 10, 15, 20 or 25 as a prediction for the Reasonableness Score. ReasonablenessScore $A$ ,  $CA$

- 1:  $Y \leftarrow 0$
- 2:  $cogScores \leftarrow []$
- 3:  $CACogScores \leftarrow$  get the cogency values for each arg. in  $CA$
- 4: **for** each argument  $C$  in  $CA$  **do**
- 5:    $DA \leftarrow$  get all the arg. components that attack  $C$
- 6:    $cogScores \leftarrow$  append the cogency values for each arg. in  $DA$
- 7:   **if**  $len(DA) > 0$  **then**
- 8:      $Y += 1$
- 9:   **end if**
- 10: **end for**
- 11:  $CV10 \leftarrow$  Count how many of the rebuttal scores in  $cogScores$  are 10
- 12:  $CV20 \leftarrow$  Count how many of the rebuttal scores in  $cogScores$  are 20
- 13:  $Q \leftarrow$  Count how many rebuttals have a cogency score lower than the counterarguments.
- 14:  $X \leftarrow$  Count how many rebuttals have a cogency score higher than all of the counterarguments.
- 15:  $Z \leftarrow$  Count how many rebuttals have a cogency score equal to the counterarguments.
- 16: **if**  $\max(cogScores) = 0$  **then**
- 17:   **return** Score 0
- 18: **end if**
- 19: **if**  $Y = 0$  **then**
- 20:   **return** Score 0
- 21: **end if**
- 22: **if**  $CV10 > \frac{len(cogScores)}{2}$  **then**
- 23:   **return** Score 10
- 24: **end if**
- 25: **if**  $1 \leq Y \leq \frac{len(CA)}{2}$  **then**
- 26:   **return** Score 10
- 27: **end if**
- 28: **if**  $Q > \frac{len(cogScores)}{2}$  **then**
- 29:   **return** Score 10
- 30: **end if**
- 31: **if**  $Y = len(CA)$  **then**
- 32:   **if**  $\max(cogScores) \geq 15$  and  $\min(cogScores) < 15$  **then**
- 33:     **if**  $\max(cogScores) > \max(CACogScores)$  **then**
- 34:       **if**  $\min(cogScores) < \min(CACogScores)$  **then**
- 35:         **return** Score 15
- 36:       **end if**
- 37:     **end if**
- 38:   **end if**
- 39:   **if**  $len(CV20) > \frac{len(cogScores)}{2}$  and  $\min(cogScores) \leq 10$  **then**
- 40:     **if**  $X > \frac{len(cogScores)}{2}$  and  $1 < Z \leq 2$  **then**
- 41:       **return** Score 20
- 42:     **end if**
- 43:   **end if**
- 44:   **if**  $\min(cogScores)=25$  and  $\max(cogScores)=25$  **then**
- 45:     **if**  $\min(cogScores) > \max(CACogScores)$  **then**
- 46:       **return** Score 25
- 47:     **end if**
- 48:   **end if**
- 49: **end if=0**

---

layer, or (iv) a SVM. Additionally, the best-performing static and dynamic embeddings were concatenated and evaluated as if they were one single embedding. The PyTorch framework [101] version 1.10 was used for implementing the LSTM model with a learning rate selected from 0.05, 0.1, RNN layers 1, 2, dropout 0.1, 0.3, 0.5, and batch size from 8, 16, 32 and a hidden size of 128. For Longformer, BERT and T5 pre-trained models, we use the PyTorch implementation of huggingface [102] version 4.16.2. For the graph, FEATHER-G embeddings, the Karate Club framework [103] was used with the standard hyperparameters. The Scikit-learn [104] framework was employed for the implementation of the Random Forest and SVM models. We trained the SVM models for each quality attribute, optimizing the Gamma and C hyperparameter (tested C range:  $10^{-4}$  to  $10^3$ , Gamma range:  $10^{-4}$  to  $10^0$ ) on the training data set given by the original split [81] in the dataset. For the rhetoric attribute, we trained the SVM models concatenating the Longformer and FEATHER-G embeddings with the emotion word embedding. The latest was obtained by (i) running the fine-tuned T5 model to detect emotions, and (ii) either using that label as an input on Glove, or extracting directly from the model the representation of the labels by summarizing the hidden states of the last four layers in the model. To train the binary classification, we converted all of the *ethos*, *pathos* and *logos* labels to *rhetorical*, while for the multi-label classification, all the non-rhetorical arguments were discarded.

Embedding	Model	f1	F1
Longformer	RandomForest	0.72	0.74
Longformer	LSTM	0.55	0.51
finetunning Longformer	dense layer	0.43	0.33
Longformer	SVM	0.74	0.72
Long. + FEATHER-G	RandomForest	0.73	0.75
Long. + FEATHER-G	SVM	<b>0.78</b>	<b>0.77</b>

TABLE 3.6: Results for the Cogency score of the 3-class sequence tagging task are given in weighted F1 (f1) and macro F1 (F1).

Embedding	Binary Clf.		Multi-label Clf		Full Pipeline	
	f1	F1	f1	F1	f1	F1
Longformer	0.78	0.69	0.70	0.62	0.91	0.57
Long.+ FEATHER-G	0.78	0.69	0.66	0.58	0.91	0.57
Long.+ FEATHER-G+ T5	0.80	0.73	0.70	0.62	0.89	0.62
Long.+ T5	0.80	0.73	0.80	0.72	0.89	0.62
Long.+ T5w/GloVe	<b>0.80</b>	<b>0.73</b>	<b>0.80</b>	<b>0.77</b>	<b>0.89</b>	<b>0.63</b>

TABLE 3.7: Results of the Argumentation Rhetoric sequence tagging task training an SVM model (weighted F1 (f1) and macro F1 (F1)).

**Results.** Table 3.6 and 3.7 report on the results for the best-performing models and embedding combinations. Performances are given on the test set in weighted average and macro multi-class F1-score. Each run was repeated five times with different random seeds to assess the stability of the results and the average score is reported. For Cogency classification, a significant improvement (from .72 to .77 macro F1-score) can be seen when the FEATHER-G graph embeddings are combined with the Longformer embeddings. The best-performing model (in bold) is composed of these embeddings along with an SVM model for the quality prediction scores.

To further probe the relationship between argument structure and quality, I conducted an additional correlation analysis. Specifically, I examined the association between argument density, defined as the number of unique argument components (claims, premises) in an essay, and the three quality dimensions. Argument density represents a purely structural measure of argumentation. We computed the Pearson correlation coefficient between density and each quality score to quantify their linear relationship. The analysis found a moderate positive correlation ( $r = 0.52$ ) between density and cogency scores in the full dataset. However, among essays with non-zero cogency, the correlation was weaker ( $r = 0.23$ ), indicating limitations of purely structural features for quality modelling. This makes sense given that by definition in the rubric by Stapleton and Wu. [19], essays with only one or no premises (i.e. low density) must receive a cogency score of 0. Only negligible to weak correlations were observed between density and reasonableness scores. These supplemental findings reinforce the importance of combining our graph-based approach with textual features to capture nuanced structural influences beyond simplistic counts.

In contrast to our work, Chuang and Yan [105] consider argument quality as only textual attributes such as acceptability and relevance, and consider argument structure as a separate aspect of a broader characterization called argumentation skill. However, accounting for these differences, we can see our results relate to theirs. They conclude that while low-quality essays tended to lack clear argument structure, there is not a strong correlation between quality and structure for essays with well-formed arguments. They further suggest argument structure and quality are related but distinct facets, with a clear structure being necessary but not fully predictive of quality. Overall, they articulate that structure and quality are associated but structure alone does not determine quality, representing overlapping yet distinct dimensions of persuasiveness and writing proficiency. Our correlation findings reinforce the limitations of pure structure metrics, aligning with and providing further confirmation of their conclusions on modelling text and structure jointly.

For Reasonableness, due to the scarceness of counterarguments and rebuttals, no deep learning model showed significant learning success. Following Algorithm 1, we obtain the Rebuttal Reasonableness score for each argument (computed starting from the cogency values obtained by our model, not the golden labels) yielding an accuracy of .80 and a macro F1 of .54 while a majority baseline obtains an accuracy of .78 and a macro F1 score of .18.

Table 3.7 shows the results for the two steps and the full pipeline of the argumentation Rhetoric classification task. The T5 fine-tuned model with Glove embeddings shows the best performance with a .73 macro F1-score for the first step of the pipeline (i.e., the binary classification *rhetoric/non-rhetoric*), a .77 macro F1-score for the multi-label classification (i.e., the multi-class classification *ethos/logos/pathos*), and a .63 macro F1-score for the full pipeline. We observe that the performance improves for every model when we add the emotion embeddings to the input feature vector, supporting our choice of integrating a general emotion dimension into the rhetorical classification for a better embedding representation. We can also notice that the graph embeddings are not really contributing to this task, leading to a detriment of macro F1-score. This result can be explained as the persuasive rhetorical strategies relies mainly on the textual formulation of the argument component itself, without being impacted by the support and attack relations involving this component.

We addressed a comparison with the state-of-the-art approaches for the Cogency assessment, despite the fact that we focus on a different dataset and divergent features (e.g., graph embeddings in our case). We retrained our model with the dataset of Wachsmuth et al. [78] following their same configuration. In this dataset of forum data, each argument instance is associated to 3 different gold labels for Cogency, one for each annotator. They also separate them into 16 different topics and train each model with 15 of them, testing on the excluded one. We followed the same process for each annotator with our baseline model (Longformer embedding + SVM). Given that they do not provide any graph structure we cannot test our best model on their data to compare. However, the results obtained are a Mean Absolute Error of .64, .38 and .52 for Expert #1, #2 and #3, respectively. Comparing with [78], we can see that for Expert #2 our baseline model outperforms their best model (.38 vs .57). In the case of Expert #1, we obtain the same result as their baseline, and for Expert #3 we perform similarly (.52 vs .50). The results we obtained on Cogency (.78 f1) are, to the best of our knowledge, the best result obtained so far in the literature [5, 78].

**Error Analysis.** A common mistake for Cogency is that the scores 0 and 25 are more often correctly classified than score 15. This is due to the imbalance of score 15 given by the nature of the essays, and the complexity of the task for human annotators, as it is easier to distinguish bad from good cogency quality, but more difficult to assess a more subtle distinction. For the argumentation Rhetoric binary classification task, the model tends to misclassify the arguments as *non rhetorical*. This results from the imbalanced dataset, where 76% of the arguments are non-rhetorical. For the multi-label classification task, the model tends to confuse pathos arguments with ethos. This can be explained by the fact that *ethos* and *pathos* are the majority and minority classes, respectively. A further extension of the dataset with the spans of text in the argument that justify the annotated rhetorical structure could yield an improvement in the performance of sequence tagging. For Reasonableness, disagreements between the results given by the algorithm and the gold labels mostly lie in a wrong

classification of the cogency score for the counterarguments and rebuttals, leading to a propagation of the error to the reasonableness score.

### 3.5 Concluding remarks

I presented a novel approach to the task of automatic quality assessment of natural language argumentation. I built a new resource of 402 students' persuasive essays annotated with 3 different quality dimensions, i.e., cogency, rhetoric and reasonableness. Through an extensive evaluation, I show that our neural architecture relying on a transformer with an attention mechanism and graph embeddings is able to successfully classify arguments along with these quality dimensions, outperforming standard baselines and similar approaches in the literature. Our quality assessment method conjugates the empirical evaluation of the cogency dimension with the graph-based computation of the reasonableness one, which encompasses the quality (expressed in terms of cogency) of the counterarguments and the argumentation structure.

In the context of AI in education, I aim to include our automatic argument quality assessment pipeline into a larger framework where the system engages the student into an explanatory rule-based dialogue to assess her essays, explain why they obtained a certain quality score and how to improve them along with the considered quality dimensions.

While this work establishes an effective approach to assessing key dimensions of argument quality, ample opportunity remains to build upon it through expanded resources and specialized modelling. For instance, larger corpora containing more balanced argumentation with extensive counterarguments and rebuttals would provide richer training data to enhance machine learning of the reasonableness dimension. Extending existing quality assessment datasets with annotated argument relations could also facilitate leveraging graphical structure. Furthermore, quality notions like cogency may require adaptation or extension to effectively evaluate arguments in specialized domains like medicine or law. Defining new principles and constructing domain-specific corpora would allow tailoring assessment to precise application needs. Overall, the presented essay evaluation framework lays the groundwork for future research to advance multi-dimensional quality modelling through richer, more diverse training data and metrics tuned to target domains.

## Chapter 4

# Integrating and Assessing External Knowledge in Medical Text

*This chapter bridges argument mining into clinical reasoning, pioneering techniques tailored for medical argumentation. More precisely, in this chapter, I introduce methods to extract clinical information from text and align it with external medical knowledge. This enables a transparent assessment of the explanatory power of different reasons based on their relevance and prevalence. By discerning between strong, simple reasons versus extraneous details, the approach ranks evidence to identify the most cogent justifications. Only highly rated reasons are then utilized to generate template-based explanations, ensuring soundness and conciseness. Moreover, by contrasting the ranked reasons against those invoked in existing explanations, the approach can critique and suggest revisions to student rationales. This promotes proper reason usage aligned with validated clinical knowledge. This chapter comprises the work published at the International Conference on Agents and Artificial Intelligence (ICAART-2023) [106] and work currently under review at the Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2023).*

Constructing high-quality explanations in the medical domain poses unique challenges in terms of reliability and transparency of the assessment process. The sensitive nature of this field demands that students thoroughly verify all premises before invocation, even if they believe a claim is true. A hallmark of medical education involves analyzing a clinical case report containing the requisite details for determining appropriate diagnoses, treatments, and next steps. Leveraging their acquired knowledge, students must correctly infer the diagnosis and provide a sound justification. Research in social sciences [107] explains the cognitive process of explanation generation. Individuals first consider all potential reasons that could explain an event. Through a process of *reason selection* [108–111], the most explanatory options are identified based on their ability to provide a sufficient yet simple account. While

these traits do not inherently co-occur, the ideal balance engenders the best explanations. For instance, when asked *why Ambrose had an accident after deviating from his routine route*, people typically cite the abnormal change, selecting from infinite possibilities via biases like abnormality detection. In medicine, symptoms exclusive to one diagnosis facilitate the elimination of alternatives, conferring significant explanatory power. With such cogent reasoning, students can formulate concise yet sound explanations, e.g., “Given symptom S, unique to diagnosis D, the patient has diagnosis D”.

This chapter investigates the automated assessment and refinement of this reason selection process. I introduce a computational methodology that inspects the clinical context, establishes connections to external knowledge in the Human Phenotype Ontology, and generates a scored list of explanatory reasons pertaining to the diagnosis. Scores reflect reasoning strength, with only highly-rated reasons used in resulting explanations. This evaluation enables the generation of basic template-based explanations which can be adopted to contrast student explanations to suggest modifications to improve soundness and sound reasoning.

## 4.1 Introduction

The limits of robustness for knowledge-dependent tasks have been thoroughly demonstrated in the literature [112]. Results show that while transformer models like BERT show some robustness to minor input perturbations, they also exhibit clear vulnerabilities. Their performance suffers on out-of-domain examples and they rely heavily on keywords and stylistic cues, rather than a deeper understanding of argument semantics. This aligns with the growing evidence that these models, trained solely on character patterns in text corpora, reach limits in their reasoning and generalization capabilities. Their knowledge remains confined to the statistics of the training data. As a result, knowledge-dependent tasks pose a particular challenge. For fine-grained relation classification or entity typing, models often lack the real-world knowledge to make inferences about connections not overtly stated in the text. In the medical domain, pertinent to this thesis, complex interrelationships are often implicitly assumed between terms and concepts. Without relevant medical knowledge, models cannot comprehend the full context.

For assessing argumentation quality, the model needs to draw inferences about the acceptability, relevance, and sufficiency of premises based on unstated background knowledge. Relying solely on contextual embeddings has clear limitations. Potential solutions involve incorporating structured knowledge into language models, such as from medical ontologies or knowledge graphs. Models like ERNIE [113, 114] integrate entity information to enrich the learned representations.

Other techniques aim to inject different types of background knowledge, not just

entities, to facilitate deeper language understanding. For example, integrating semantic role information [115] can provide models with knowledge about predicate-argument structures to better capture relational reasoning. Commonsense knowledge graphs like ConceptNet [116] have also been utilized to inject conceptual relationships between everyday events and concepts. This can aid in implicit causal reasoning and drawing inferences. Models like KI-BERT [117], CALM [118], and DiffKG [119] take the approach of injecting external knowledge graphs to teach language models relational knowledge and reasoning capabilities. KI-BERT injects entity and concept knowledge from sources like ConceptNet and WordNet to improve semantic understanding. CALM uses a self-supervised objective to infuse commonsense conceptual knowledge from text into T5 models. DiffKG incorporates interpretable knowledge graph reasoning directly into transformer models for dialogue systems. By enabling models to perform multi-hop reasoning on knowledge graphs, these approaches equip language understanding models with structured factual, conceptual, and commonsense knowledge beyond what is contained in text corpora.

In the medical domain, medical ontologies such as UMLS [49] or HPO [50] can be leveraged to incorporate hierarchical medical concepts and terminology. This domain knowledge teaches models the nuanced differences between related medical terms.

These approaches demonstrate that providing models with external structured knowledge beyond surface patterns enables more robust relational and inferential understanding. This allows for tackling knowledge-dependent tasks where vital contextual associations are not explicitly stated. In argument assessment, integrating conceptual knowledge from sources like medical ontologies can enable models to make judgments requiring domain expertise. Rather than relying on fragile distributional statistics, structured knowledge infusion equips models with semantics, common sense, and reasoning capabilities crucial for reliable and generalizable argument understanding.

While knowledge injection techniques show promise for making models more robust, a key challenge lies in validating whether the knowledge is properly integrated and applied. For sensitive applications like healthcare, it is critical to ensure models utilize knowledge appropriately when making high-stakes decisions. This motivates the need for explainable AI techniques that can evaluate how well the model reasoning aligns with established domain knowledge. Explainability enables auditing the model's inferences to verify sound and transparent usage of the injected knowledge.

Explainable artificial intelligence (XAI) has become a significant research focus given the prevalence of opaque models and their use in sensitive domains like medicine [120, 121]. Although AI systems aid decision-making, their efficacy depends on providing comprehensible, useful explanations to users [122, 123]. However, leading XAI techniques often yield unsound or redundant explanations [124]. This chapter puts forth a methodology to automatically assess explanation quality in medicine,



ensuring transparency by judging the relevance of reasons against a standardized ontology. By extracting clinical details from clinical cases and mapping them to validated concepts, the approach can identify strong explanatory reasons grounded in medical knowledge. Explanations only utilizing highly-rated reasons based on statistical prevalence are deemed higher quality. Moreover, contrasting system-generated explanations with student rationales allows critiquing and suggesting improved reasoning better aligned with the knowledge. This promotes proper utilization of domain expertise.

Specifically, we enable generating natural language argument-based explanations for medical resident training. In these exams, residents analyze a clinical case and select the correct diagnosis from several options. They must also justify their choice. We automate explaining why one diagnosis is correct and others incorrect through arguments based on relevant case symptoms. Additionally, our approach contrasts student explanations to recommend modifications enhancing soundness and simplicity.

First, a pipeline extracts and matches symptoms from cases to a medical knowledge base [125], identifying associated diseases and frequencies. Then, explanatory patterns employ this data to generate arguments. We annotated 314 clinical cases and evaluated this approach, achieving promising results. This work addresses the lack of linguistically annotated medical resources for explanation generation.

Subsequently, we score explanation reasons via a prevalence function and external knowledge, judging pertinence transparently. Applied to medicine, our technique evaluates student explanations against computed relevance ratings. It leverages the Human Phenotype Ontology (HPO) [50] and a deterministic prevalence function to rate reasons based on a case's context. Analysis of 621 expert-explained cases demonstrates the method's effectiveness.

Though assessed on medicine, this methodology extends to any explainable domain given appropriate knowledge bases and prevalence functions. It assists resident training by evaluating explanation reasoning. For online discussions, it helps identify high-quality explanations, promoting informed dialogue. Overall, this research enables a systematic, transparent assessment of explanation reasoning, especially for medicine.

The chapter is organized as follows. Section 4.2 covers related work. Section 4.3 presents the clinical information extraction and matching pipeline. Section 4.4 details assessing explanatory power through scoring via the prevalence function. Section 4.5 generates natural language explanations from matched symptoms. Section 4.6 concludes and discusses applications to education and online discussions.

## **4.2 Related Work**

**Clinical NLP techniques** Since the introduction of BERT [21], transformer-based models have recently had a major impact on most NLP tasks. Multiple models evolved from it with different design choices, like RoBERTa [126], ELECTRA [127] and ALBERT [70]. These models are trained on a large amount of data from multiple sources and domains, which means that they are not necessarily prepared for the biomedical domain.

In recent years, a great number of resources and NLP tools have been developed specifically for the biomedical domain. For entity extraction, the most popular datasets are BC4CHEMD [9], B5CDR-Chem [10], NCBI-Disease [11], BC2GM [12], JNLPBA [13], where the annotations range from drug-disease interactions to the identification of diseases, genes, and molecular entities such as protein, DNA, RNA. Symptom detection, i.e., the task we address in this chapter, can be seen as a sub-task of the broader task of medical entity extraction.

Off-the-shelf NLP tool-kits such as Spacy [128], MedSpacy [129] and CLAMP [130] provide multiple modules for text processing. In particular, MedSpacy is built on top of Spacy specifically for clinical natural language processing, while CLAMP offers a method for named entity recognition (NER) as well as a visual interface for annotating and labeling clinical text.

Most of the recent approaches treat NER as a sequence labeling task where *specialized* transformer-based models hold the best results. For example, [131] showed that pre-training the ALBERT model on a huge biomedical corpus ensured that the model captured better biomedical context-dependent NER. Results outperform non-specialized models obtaining SOTA results in a lot of datasets. Similar results can be seen in [132], where the authors pre-train a biomedical language model using biomedical text and vocabulary with the technique proposed by ELECTRA. Other specialized models based on BERT have been proposed by [133], [134] and [68] and BioMed-RoBERTa [135] based on RoBERTa.

[136] propose UmlsBERT, a contextual embedding model that integrates domain knowledge from the Unified Medical Language System (UMLS) [49], taking into consideration structured expert domain knowledge. They show that UmlsBERT can associate different clinical terms with similar meanings in the UMLS knowledge base and create meaningful input embeddings by leveraging information from the semantic type of each word. In our work, we compare the representation of the symptoms found in the clinical case with different contextual embeddings with the goal to find a representation which matches the one provided in the Human Phenotype Ontology (HPO).

Ngai et al. [137] also tackle the problem of finding relevant clinical information, where among the entities they also identify symptoms. In contrast to our work, they only focus on 6 specific diagnoses. Furthermore, their goal is to predict the correct diagnosis and explain these predictions using feature attribution methods, whilst ours is to generate high-quality explanations in natural language for educational purposes, i.e., to improve medical residents' skills in explaining their answers to the

exams.

Besides detecting symptoms from clinical cases, in our work, we also aim to accurately map them to medical ontologies, such as the Human Phenotype Ontology (HPO), to identify the relationship between the symptoms (originally described in layperson terms) and diseases. Recent work by [138] proposes a tool for automatically translating between layperson terminology and HPO, using a vector space and a neural network to create vector representations of medical terms and compare them to layperson versions. However, this approach has a limitation in that it translates layperson terms without considering their context, potentially missing relevant information that may change the semantics of the term. In our work, we propose a method that takes into account the context in which the layperson term is introduced, leading therefore to an accurate mapping to an HPO term.

**Natural Language Explanation Generation** Natural language explanation generation has received a lot of attention in recent years, grounding on the progress of generative models to train specific models for explanations. [139] generate explanations by justifying a relation (*entailment*, *contradiction* or *neutral*) for a premise-hypothesis pair by training a Bi-LSTM on their e-SNLI dataset, i.e., the Stanford Natural Language Inference [140] dataset augmented with an explanation layer which explains the SNLI relations. [141] propose to generate short explanations with GPT-2 [142], learned together with the input by a classifier to improve the final label prediction, using e-SNLI [139]. These solutions are not applicable to our use case given that explaining a medical diagnosis is a more challenging task than restraining the explanations to the three basic relations considered by [139] and [141]. [143] propose an approach based on the T5 model [144] to generate an explanation after prediction. Again, this solution is not applicable to the specific medical scenario we target, where explanations require to be structured following precise argumentative structures [145–147] and to ground on medical knowledge, like the one we inject through the HPO.

Other approaches use explanations via templates [148], e.g., [149] uses templates and inject the reasoning steps and query of their Q&A system. To the best of our knowledge, no related work generates natural language post-hoc explanations for the medical domain.

**Explanation Selection.** While the previous sections covered approaches related to extracting and aligning clinical information, assessing the explanatory power of reasons requires modelling how humans select explanatory causes. The cognitive science literature provides theoretical frameworks on how people determine the most relevant reasons to explain events. The following subsection reviews research on explanation selection in psychology that informs our methodology for discriminating between explanatory clinical reasons.

In the process of explanation selection, individuals choose what they perceive to be the most relevant causes from a larger set of causes for a particular event. This selection is not arbitrary and is guided by criteria such as temporality, abnormality, intention, and the differences between a fact and a foil [107]. Hilton [150] sustains this is due to the fact that causal chains are often too large to comprehend. Research shows that the primary way individuals select explanations is by contrasting a fact and a foil. The fact refers to the actual state of affairs, while the foil represents an alternative state that did not occur. The contrast between the fact and the foil forms the basis for explanation selection, with the explanation that highlights the greatest number of differences between the fact and the foil deemed to have the highest explanatory power [151]. Contrastive explanation is a concept that further elaborates on this idea. It posits that the differences between two events form the basis for explanation. This theory has garnered support from experimental research in cognitive science, which suggests that people perform causal inference, explanation, and generalization based on contrastive cases [152, 153].

Abnormality also plays a crucial role in explanation selection. Hilton and Slugoski [111] propose the abnormal conditions model, arguing that abnormal events are key in causal explanation. This model suggests that individuals use their perceived background knowledge to select conditions that are considered abnormal. This model has been supported by subsequent experimental studies [154–156]. In this chapter, we introduce an approach that not only evaluates the relevance of each potential explanation for a given event but also incorporates the principles of abnormality and contrastive explanation into the calculation of the relevance score.

**XAI for the Medical Domain.** The importance of explanations in AI systems, particularly in the medical domain, has been extensively studied [157–159]. In the context of medical diagnosis, explanations often involve identifying the key reasons or symptoms that led to a specific diagnosis. The Human Phenotype Ontology (HPO) [50] provides a standardized vocabulary of phenotypic abnormalities encountered in human disease, which can be used to facilitate the assessment of explanations in this domain. Our work builds upon this ontology by developing an approach that assesses the selected reasons in explanations. The National Institutes of Health (NIH) Undiagnosed Diseases Program (UDP) [160] has also investigated the use of HPO in the context of diagnosing and evaluating patients with conditions that have eluded diagnosis. The clinical features of a patient are encoded into HPO terms, which are then used to retrieve a list of candidate diseases that might explain the patient’s phenotype. This list is then examined by a clinician to identify the most likely diagnosis. Our methodology extends this approach by not only using HPO to facilitate diagnosis but also to evaluate the reasons given in explanations.

### 4.3 Extracting and Aligning Clinical Information

The proposed approach relies on specialized techniques to identify relevant clinical details from the text and map them to validated medical knowledge. This section describes the dataset, methods, and experiments conducted for this extraction and alignment pipeline. The clinical case dataset used to develop and evaluate these methods is first outlined, including statistics and annotation details. Then, an overview is provided of the pipeline architecture and its key steps - symptom extraction using neural models and alignment to ontology terms based on contextual embeddings. Experiments demonstrate promising results, with a top performance of 0.86 F1 for entity recognition and 0.53 accuracy for symptom matching. Error analysis reveals challenges like mental health diagnoses lacking clear ontology linkage. Overall, this section delineates empirical work enabling the extraction and grounding of explanatory clinical evidence, forming a critical foundation for subsequent relevance assessment.

#### 4.3.1 Dataset

To train and evaluate the proposed approach to build natural language explanatory arguments, we rely on the MEDQA dataset [161], which contains a set of clinical case descriptions together with a set of possible questions and answers on the correct diagnosis. The questions and their associated answers were collected from the National Medical Board Examination in the USA (USMLE), Mainland China (MCMLE), and Taiwan (TWMLE). In this work, we only focus on the clinical cases and the questions in English (i.e., USMLE). In total, the MEDQA-USMLE dataset consists of 12,723 unique questions on different topics, ranging from questions like “Which of the following symptoms belongs to schizophrenia?” to questions about the most probable diagnosis, treatment or outcomes for a certain clinical case which is described [161]. To reach our goal, we extract the clinical cases belonging to the latter group, which are intended to test medical residents to make the correct diagnosis. We end up with 314 unique clinical cases associated with the list of possible diagnoses.

**Annotation of the MEDQA-USMLE Clinical Cases.** To annotate the clinical cases from the MEDQA-USMLE dataset, we rely on the labels from the Unified Medical Language System (UMLS) [49] Semantic Types, making it consistent with standard textual annotations in the medical domain [162–164]. In particular, we annotate the following elements in the clinical case descriptions: *Sign or Symptom*, *Finding*, *No Symptom Occurrence*, *Population Group*, *Age Group*, *Location* and *Temporal Concept*. In this work, we use only the symptoms, but we addressed a complete annotation to employ these data for future work. Quantifiers defining a symptom have not been annotated (e.g., we can find “moderate pain”, where we only annotate “pain”). The labels *Sign or Symptom* and *No Symptom Occurrence* are associated only to the

text snippet defining the symptom in a sentence. *Findings* consist of such information discovered by direct observation or measurement of an organism’s attribute or condition. For instance, *components* in "Her temperature is 39.3°C (102.8°F), pulse is 104/min, respirations are 24/min, and blood pressure is 135/88 mm Hg". *Location* refers to the location of a symptom in the human body, and *Temporal Concept* is used to tag time-related information, including duration and time intervals. *Population Group* and *Age Group* highlight information on the age and gender of the patient.

To address the annotation process of the MEDQA-USMLE dataset, we first carried out a semi-automatic annotation relying on the UMLS database [49]. We processed each clinical case through the UMLS database and obtained all the entities detected along their Concept Unique Identifiers (CUI) and their semantic type. The semantic type is then used to disambiguate the entities and generate the pre-annotated files. After the definition of the detailed annotation guidelines (summarized above) in collaboration with clinical doctors, three annotators with a background in computational linguistics carried out the annotation of the 314 clinical cases.

To ensure the reliability of the annotation task, the inter-annotator agreement (IAA) has been calculated on an unseen shared subset of 10 clinical cases annotated by four annotators, obtaining a Fleiss’ kappa [165] of 0.70 for all of the annotated labels, 0.61 for *Sign or Symptom*, 0.94 for *Location*, 0.71 for *Population Group*, 0.66 for *Finding*, 0.96 for *Age Group* and 0.96 for *No Symptoms Occurrence*. We can see a substantial agreement for *Sign or Symptom*, *Finding* and *Population Group*, and an almost perfect agreement for *Location*, *Age Group* and *No Symptoms Occurrence*.

Table 4.1 reports on the statistics of the final dataset, named MEDQA-USMLE-Symp.<sup>1</sup> The accuracy of the annotations provided by the three annotators has been validated from a medical perspective with a clinical doctor. Of the seven entity labels, only three contain medical vocabulary (*Sign or Symptom*, *Finding*, and *No Symptom Occurrence*) and they have been evaluated by this expert. More specifically, we randomly sampled 10% of the data (i.e., 30 cases) and we asked the clinician to verify whether the entity was correctly labeled and whether there were any missing or extra words. The results of the validation showed that 98% of the data was labeled correctly. Less than 2% of the instances were evaluated as incorrectly labeled (e.g., a *Finding* that was labeled as a *Sign or Symptom* or vice versa).

TABLE 4.1: Statistics of the MEDQA-USMLE-Symp dataset.

Label	# Entities
Sign or Symptom	1579
Finding	1169
Temporal Concept	567
Location	498
Population Group	364
Age Group	304
No Symptom Occurrence	264

<sup>1</sup><https://github.com/Wimmics/MEDQA-USMLE-Symp>

**External Knowledge of Diseases and Relevant Symptoms.** To collect the medical knowledge needed to define whether a detected symptom is relevant with respect to a given disease, we employ the HPO knowledge base to retrieve (i) the relevant information of each diagnosis which is proposed as an option to answer the question "Which of the following is the most likely diagnosis?", and (ii) the symptoms (named *terms* in HPO) associated to each diagnosis. This knowledge base also includes information on the frequency<sup>2</sup> of the occurrence of symptoms, defined in collaboration with ORPHA<sup>3</sup> as follows: Excluded (0%); Very rare (1-4%); Occasional (5-29%); Frequent (30-79%); Very frequent (99-80%). Obligate (100%); HPO integrates different sources of symptoms, including ORPHA and OMIM<sup>4</sup>. This knowledge base is quite rich and contains also links and hierarchical links between symptoms (Symptom S2 subclass of Symptom S1), genes or definitions.

### 4.3.2 Proposed framework

An overview of the framework we propose to address automatic symptom relevancy assessment and matching to build our natural language explanations is visualized in Figure 4.1. Starting from the clinical cases in which the correct and incorrect diagnosis are already identified, the goal is to assess the relevant symptoms present in the case such that these symptoms can be used to *explain why* a certain diagnosis is the correct one and *why* the incorrect ones have to be discarded.

In order to accurately diagnose a patient's condition, it is important to identify the symptoms that are most relevant to the possible diagnoses. This means looking at all of the symptoms that have been detected and determining which ones are most likely to be related to the underlying cause of the patient's condition. This can be done by considering the individual symptoms and their potential connections to the possible diagnoses. It is also important to consider any additional information that may be available, such as the patient's medical history and other relevant factors, in order to be able to fully explain the diagnosis. Our work focuses on identifying relevant symptoms in order to accurately diagnose a patient's condition.

The relevancy assessment model associates, when possible, the pertinent symptoms mentioned in the clinical case description with a symptom of a diagnosis found in the HPO knowledge base. The proposed framework consists of two different steps, where: (i) we retrieve from HPO the required diagnosis information (i.e., the symptoms and how common they are), then the symptoms in the case are detected and extracted using an attention-based neural architecture which relies on the clinical case text only; (ii) the relevancy of each symptom is assessed by matching the detected symptoms with the ones retrieved from HPO. The matched symptoms

<sup>2</sup><https://hpo.jax.org/app/browse/term/HP:0040279>

<sup>3</sup><https://www.orpha.net/consor/cgi-bin/index.php?lng=FR>

<sup>4</sup><https://www.omim.org/>

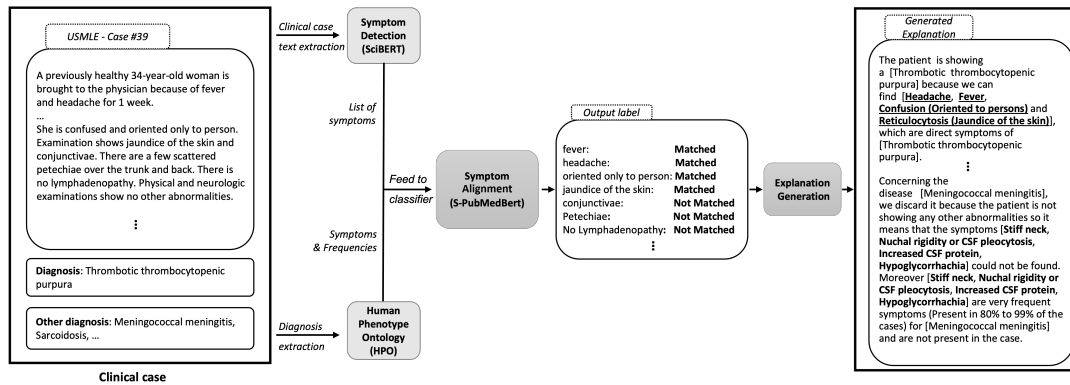


FIGURE 4.1: Overview of our full pipeline for symptom prediction and alignment, and NL explanation generation module.

are then used to generate natural language argument-based explanations for correct and incorrect diagnoses. In the following, we explain in detail each sub-task in the pipeline:

**Symptoms detection**, consisting in detecting the different symptoms described in the clinical case (medical terms or symptoms described by the patient’s own words). In order to detect these entities, we propose a neural approach based on pre-trained Transformer Language Models.

**Symptoms alignment**, to align a symptom detected in the clinical case with an identical term in HPO. We first compute an embedding vector for each found symptom and then calculate the cosine distance with each term in HPO. We then assign the closest concept to that symptom. We evaluated both static and contextual embedding methods.

**Explanation generation** We propose template-based explanations based solely on the symptoms that are relevant to explain the diagnosis. To do this we propose several templates that tackle different kinds of explanations, going from explaining why a patient was given a certain diagnosis to explaining why the alternatives cannot be considered viable options. We support our explanations with statistical information obtained from HPO such as the frequency of each symptom incidence, and we propose to look for possible symptoms that were not detected by the system but are frequent for a certain disease.

### 4.3.3 Evaluation

In this section, we report on the experimental setup, the obtained results and the error analysis for the symptom detection and symptom alignment methods.

**Setup.** For the symptom detection task, we experimented with different transformer-based Language Models (LMs) such as BERT [21], SciBERT [133], BioBERT [134], PubMedBERT [68] and UmlsBERT [136] initialized with their respective pre-trained weights. All the models we employ are specialized in the biomedical domain, with



the exception of BERT which will serve us as a baseline. We cast the symptom detection problem as a sequence tagging task. Following the BIO-tagging scheme, each token is labeled as either being at the **B**eginning, **I**nside or **O**utside of a component. This translates into a sequence tagging problem with three labels, i.e., *B-Sign-or-Symptom*, *I-Sign-or-Symptom* and *Outside*. To fine-tune the LMs, we use the PyTorch implementation of huggingface [102] (v4.18). For BERT, we use the uncased base model with 12 transformer blocks, a hidden size of 768, 12 attention heads, and a learning rate of  $2.5e-5$  with Adam optimizer for 3 epochs. The same configuration was used to fine-tune SciBERT BioBERT, PubMedBERT and UmlsBERT. For SciBERT, we use both the cased and uncased versions, and for BioBERT we use version 1.2. Batch size was 8 with a maximum sequence length of 128 subword tokens per input example.

Regarding the matching module, we experimented with two different methods to align our detected symptoms with terms in HPO by (i) directly comparing the computed embeddings of the detected symptoms with the embeddings of the terms in HPO, and (ii) by taking into account the context in which the symptoms are detected and applying the same context to every term in HPO.

To align our detected symptoms (in the clinical case) with the equivalent HPO terms, we calculate the cosine distance of each embedding of the HPO terms with respect to the embedding of the detected symptom. In the experimental setting of (i) and (ii), we use the static pre-trained embeddings GloVe 6B as well as BERT, SciBERT, BioBERT and UmlsBERT in the same configurations as in the symptom detection task. For (ii), it is necessary to calculate the context embeddings "on the fly" because each context is unique and depends on the clinical case where it was detected. It is not reasonable to recalculate all HPO term embeddings on the fly for each new context since the ontology contains 10,319 unique terms, so we propose to generate all the HPO terms embedding at once and save them. Therefore, this module takes as input the symptoms detected by the previous module and finds the context<sup>5</sup> of these symptoms in the clinical case.

The context  $C$  is embedded using sentence embedding methods and saved separately from the symptom  $S$ , and the two embeddings are added together ( $C + S$ ) to form the reference  $R$ . This same context embedding  $C$  is added in the same way to each HPO term embedding  $T_1, T_2, \dots, T_i$  to form the candidates  $C_1, C_2, \dots, C_i$ .

We compute and retrieve the five best cosine distances between  $C$  and  $R$  to address a fair comparison with other systems.

We defined a test set of 23 cases where (i) we retrieved from HPO the symptoms related to the diseases for each case, and (ii) we manually aligned the annotated symptoms in the case to the concepts from HPO. This resulted in 162 symptoms aligned to a specific term in HPO that serve us as a testing set for our matching module.

<sup>5</sup>The context consists of the sentence(s) containing the symptom and the entire clinical case.

As mentioned in Section 4.2, the system proposed by [138] offers a similar approach to translating layperson terms to medical terms in HPO. However, their work does not take into account the context in which a symptom is found. To the best of our knowledge, this system constitutes the state-of-the-art when translating layperson terms to HPO terms so we decided to compare our proposal with theirs. However, due to the unavailability of their model, we rely on their online demo, which outputs only the top 5 ranking of the HPO terms that are closest to the input symptom. To perform a comparison with our pipeline, we first compute the accuracy of the aligned symptoms using our symptoms alignment module and then replace it with [138] proposed system (DASH). Results are shown in Table 4.4.

Since a symptom can be composed of several words (e.g., "shortness of breath"), we split the symptom into words that we encode by either using each word as an input on Glove [58], or extracting directly from the contextualized models the representation of the symptom by summarizing the hidden states of the last four layers in the model. We then sum the vectors of each word to get an n-gram representation of the symptom. We also explore sentence embeddings, by making use of Sentence-BERT [71], a new model that derives semantically meaningful sentence embeddings (i.e., semantically similar sentences are close in vector space) that can be compared using cosine similarity. Sentence-BERT can be used with different pre-trained models, in this work we focus on the models BERT [21], SciBERT [133], UMLSBERT [136] and S-PubMedBert by [166]. The first represents a competitive baseline in our experiments since it is the SOTA model for comparing sentences cross-domain, while the three latter models are pre-trained on scientific or medical data or both.

To tackle both tasks we make use of our annotated dataset (Section 3). The annotations are converted into two datasets, one for each part of the pipeline. The first dataset is used for the symptom detection task, and it is in the CoNLL format for token-wise labels. The second dataset, for the symptom alignment task, is converted into a csv format, where each symptom in the clinical case description and available related knowledge (i.e., the list of symptoms and their frequencies for each possible diagnosis associated with the case) extracted from HPO are paired.

**Results.** Results for the symptom detection task are shown in Table 4.2 in macro multi-class precision, recall, and F1 score. We can observe that all models perform similarly, with the best results from the specialized SciBERT [133] model. The biggest difference in performance is given by comparing SciBERT uncased with PubMedBERT, with the SciBERT model performing better. Interestingly, BERT performs closely to the specialized models, and, in some cases, it outperforms them. This may be due to the fact that the clinical cases from our dataset are written for medical exams at the med school. They contain some technical specialized words, but overall the symptoms are described in layperson terms.

It is worth noticing that the majority of our labels do not pertain to medical terminology (e.g. Age and Population Group, Location and Temporal Concept). Sign

TABLE 4.2: Results for entity recognition in macro multi-class precision, recall, and F1-score.

Model	P	R	F1
BERT	0.85	0.84	0.84
BioBERT v1.2	0.84	0.85	0.84
UmlsBERT	0.85	0.85	0.85
PubMedBERTbase	0.83	0.84	0.83
SciBERT cased	0.85	0.85	0.85
SciBERT uncased	<b>0.85</b>	<b>0.86</b>	<b>0.86</b>

TABLE 4.3: Results for entity recognition using our best performing model (SciBERT uncased) in P, R, and F1-score.

Entity	P	R	F1
Other	0.93	0.91	0.92
Age Group	1.00	0.97	0.98
Finding	0.85	0.88	0.86
Location	0.74	0.80	0.77
No Symptom Occurrence	0.79	0.72	0.75
Population Group	0.88	0.95	0.91
Sign or Symptom	0.83	0.82	0.82
Temporal Concept	0.78	0.87	0.82
Weighted avg	0.89	0.89	0.89
Macro avg	0.85	0.86	0.86

or Symptom and Finding are the only labels that require specialized vocabulary.

Overall, SciBERT uncased is the best-performing model (in bold) with a macro F1-score of 0.86, outperforming the other approaches for each of the categories. In Table 4.3 we report the performances for each entity with the best-performing model. The *Sign or Symptom* detection task obtains a 0.82 F1 score. In the work of [137], the authors also detect symptoms obtaining an F1 score of 0.61. However, these results can not be directly compared since the datasets on which both models were fine-tuned are different: we train on clinical cases, while they use dialogues between doctors and patients. Moreover, given that the dataset they use is not released, we can not evaluate our approach to their data.

The results of the symptoms alignment module experiments are summarised in Table 4.4. As baseline models, we propose to use the same methods but without the context of the symptom, similarly to [138] *DASH*. In Table 4.4 we show only the best-performing baseline *PubMedBERT no context* obtaining similar results to *DASH* (0.41 and 0.37, respectively). Adding contextual representation to the embeddings results in a significant improvement (up to 0.53 in accuracy) supporting the hypothesis that context plays an important role when translating layperson terms to formal medical terms.

**Error Analysis.** HPO has limitations with respect to the number of symptoms associated with each diagnosis. For some diagnoses, we have multiple symptoms, while

TABLE 4.4: Results for DASH and our symptom alignment method using different embeddings with and without context (accuracy score).

<b>Model</b>	<b>Accuracy</b>
DASH	0.37
PubMedBERT no context	0.41
BERT + context	0.38
SciBERT + context	0.39
UMLSBERT + context	0.44
S-PubMedBERT + context	<b>0.53</b>

for others we can have only one or none. We have observed that the model tends to make more mistakes when diagnosing mental diseases. Upon further inspection of the nature of HPO for such diagnoses, we have found that either the diagnosis is not present in the HPO ontology, or the symptoms listed are more general in nature. These symptoms include common ones such as *changes in appetite* or *low energy*, which alone may not be relevant, but when considered together, may indicate a precise diagnosis. Additionally, some relevant symptoms may not be explicitly described but encoded in the clinical cases as *Findings*.

These findings often refer to a relevant symptom that is not explicitly mentioned in the case, like in the example introduced in Section 3 about findings, where we have "respirations are 24/min" that, combined with the fact that the patient is a 34-year-old woman, means that she has *dyspnea*. Automatically deriving this implicit knowledge remains an open challenging issue. Given that we rely on HPO only, some diseases or diagnoses are not present in the knowledge base, preventing us from generating the associated explanations. Combining HPO with more specialized medical knowledge bases is a future direction for this work, both to complete the information we have and also to integrate new diagnoses.

## 4.4 Assessing Explanatory Power

In domains like healthcare and education, the ability of an AI system to explain its reasoning is critical for acceptance and effective use. However, recent research reveals that many state-of-the-art explainable AI (XAI) [120, 121] techniques fail to generate high-quality explanations that users can intuitively comprehend and validate [122, 123]. Explanations often lack sound justifications, exhibit redundancy, or do not match how humans perform explanation selection and assessment [124]. This demonstrates an open challenge in developing AI systems capable of producing cogent, non-redundant explanations that manifest transparent rationale.

A crucial step towards explainable AI is the ability to automatically assess the quality of explanations by evaluating the explanatory power of the underlying reasons. Explanatory power refers to how strongly a reason influences the occurrence of an event, based on relevance, abnormality, and contrast to alternatives [111, 151].

By algorithmically scoring explanation reasons on these key criteria, we can determine the strength of the provided justifications. This facilitates comparison to human explanation best practices and enables iterative improvements to explanation generation. In our context, medical students play the role of explainers, where they have to provide explanations for different clinical cases.

Building on the ability to identify and encode relevant clinical features into standard terminology, the next step is determining which of these features serve as the strongest reasons to explain a given diagnosis. To enable such assessment in a systematic and transparent manner, we propose a novel approach that leverages an external knowledge base and a deterministic prevalence function. This prevalence function acts similarly to a recipe, taking into account different predefined conditions (like relevance, abnormality, and contrast) and returning a pertinence score in a transparent, step-by-step manner.

When applied to the medical field, our approach scrutinizes explanations provided in medical examinations, wherein medical residents elucidate a specific diagnosis of a patient, given the context (i.e., a clinical case detailing the patient's condition) and their medical expertise. Consequently, we generate an assessment that identifies the reasons employed in the explanation and evaluates them against the relevance scoring produced by our approach. Our approach leverages an external knowledge base, the Human Phenotype Ontology (HPO) [50], and a deterministic prevalence function to score each reason based on its pertinence in the domain. This function allows to elaborate the resulting reasons' scores, in a transparent way. We evaluate our approach on the Antidote Casimedicos dataset [167], a unique resource comprising 621 clinical case descriptions, each with a set of potential diagnoses, an indicator of the correct answer, and a detailed explanation of the decision-making process provided by medical professionals. The results obtained on this dataset show the effectiveness of the proposed approach.

While our methodology is assessed on a use case from the medical domain, it is abstract enough to be applied to any domain. We envision two potential scenarios where our methodology could be particularly beneficial: AI for education and online medical fora. In the context of AI for education, our approach can assist medical resident students in learning how to solve medical cases. By providing a systematic and transparent way of evaluating the reasons given in explanations, we can help students to understand the rationale behind a specific diagnosis and why other potential diagnoses are dismissed. In online medical fora, our approach can help online users to distinguish good explanations from bad explanations in medical fora. Users often discuss diagnoses and share their experiences, but the quality of these discussions can vary widely. With our approach, we can provide a systematic and transparent way of evaluating the reasons given in these discussions, helping users and moderators identify high-quality explanations and promote more informed discussions.

The research presented in this chapter is driven by the necessity for a systematic

and transparent methodology to assess the pertinence of reasons used in medical explanations. To the best of our knowledge, this is the first approach that leverages an external knowledge base, the Human Phenotype Ontology (HPO), and a deterministic prevalence function to evaluate the reasons of the potential diagnoses based on their relevance in the context of a specific clinical case and grounding on the HPO knowledge base.

#### 4.4.1 Assessing Reasons used in Explanations

My approach to assessing the reasons used in explanations is visualized in Figure 4.2. I start with a clinical case of a patient, supplemented by an explanation provided by a medical expert, which elaborates on the specific diagnosis attributed to the patient. The objective is to evaluate the reasons invoked by the expert to justify the medical diagnosis. To achieve this, I compute a *pertinence score* using a deterministic prevalence function (see Section 4.4.4), which ensures complete transparency, allowing us to explain why each reason is more or less pertinent than the others with respect to the given case.

The HPO Ontology [50] serves as our external knowledge base (KB), providing a standardized vocabulary of phenotypic abnormalities encountered in human diseases. This external KB is used to facilitate the evaluation of the reasons given in the explanation.

The approach I propose consists of two main steps: (i) the reasons given in the explanation are extracted from the clinical case and encoded into HPO terms (following the approach proposed in Section 4.3), in which a medical Named Entity Recognition (NER) step is performed, to then align them into HPO terms. This allows us to retrieve all the standardized information the ontology contains, such as the definition and the occurrence rate of that term in each possible disease; (ii) the pertinence score for each reason is computed using the prevalence function, which takes into account the relevance of each reason in the context of the specific clinical case and the knowledge base.

#### 4.4.2 Data Preprocessing

While the MEDQA-USMLE-Symp dataset 4.3.1 allowed us to identify and encode relevant clinical features, our next goal is assessing the explanatory power of those features in expert diagnoses. For this, we need a dataset that provides expert explanations justifying the reasoning behind each diagnosis. To obtain such data, we utilize the Antidote Casimedicos dataset [167], which shares similarities with MEDQA-USMLE-Symp in its medical domain and exam-style clinical case structure. It contains 621 clinical cases describing patient symptoms and history, and like MEDQA-USMLE-Symp, each case has a set of possible diagnoses and an indicator of the correct answer.

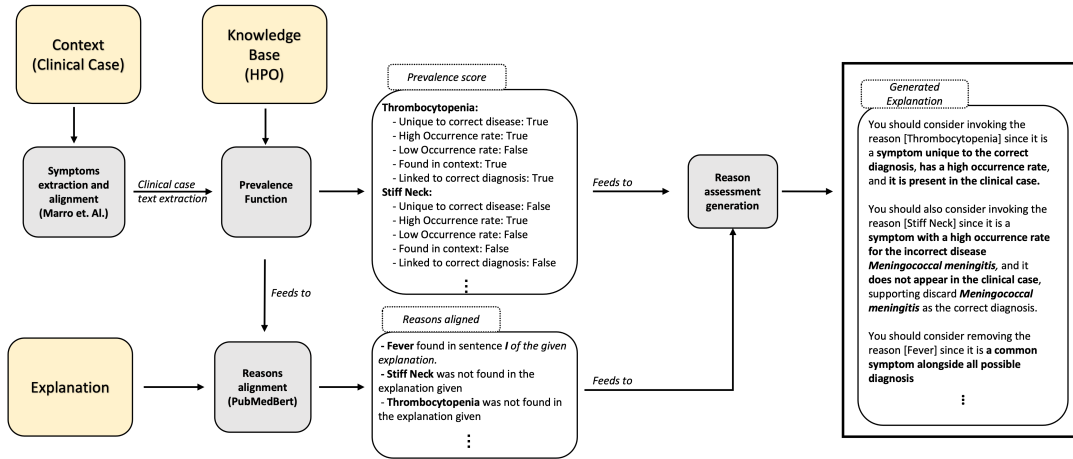


FIGURE 4.2: Overview of our approach for the automatic assessment of explanation's reasons.

The key difference is that the Antidote Casimedicos dataset also provides detailed explanations written by medical professionals, justifying their diagnosis and contrasting alternatives. These explanations offer a rich basis for evaluating how doctors select the most pertinent reasons to explain diagnoses.

Since the two datasets share a similar clinical case structure, the natural language processing techniques used previously to extract symptoms from case descriptions remain applicable. The Antidote Casimedicos data builds on the encoded clinical features by adding expert rationale. This allows us to assess feature relevance against doctor-provided explanations, enabling explanation quality evaluation.

To prepare the data, we first enhance contextual information in the explanations by expanding abbreviated diagnosis references. It is common for the explainer to refer to diagnoses as “Answer 1”, thus we implement a string replacement with the corresponding answer. Subsequently, as delineated in [167], the dataset encompasses various types of questions. For the purpose of evaluating the explanations provided by the experts, we manually filter out cases that solely discuss potential diagnoses of the patients, yielding a total of 206 clinical cases.

#### 4.4.3 Identification and Alignment of Potential Causes

The initial phase of our approach (Figure 4.2) consists in identifying all potential causes within the given context. In our medical scenario, this context is represented by clinical cases, and we regard all symptoms as potential causes that could explain the patient's diagnosis. To address this, we employ the approach proposed in Section 4.3, which performs two critical steps: the first step involves the recognition of medical-named entities within the text, and the second step aligns these identified entities with the corresponding terms from HPO.

#### 4.4.4 Prevalence Function

The Prevalence Function 2 is a central component of our approach, designed to systematically assess the pertinence of each possible reason that could explain a given event. This function is inspired by cognitive processes involved in explanation selection, aiming to replicate these processes in a transparent and replicable manner. The function takes into account the possible reasons found in the clinical case (see Section 4.4.3) and a KB that serves to find other possible reasons outside of the context but still relevant to the case. It then evaluates all possible reasons based on a set of predefined conditions, each of which contributes to the final prevalence score of the key reason. These conditions include whether the key reason is linked to the correct or incorrect diagnosis, its occurrence rate, and whether it is unique to the correct diagnosis, or shared among all possible diagnoses. In line with the abnormal conditions model, our function assigns a higher score to key reasons that are unique to the correct diagnosis and have a low occurrence rate. This reflects the idea that abnormal conditions, i.e., conditions that do not usually occur, are more likely to be the cause of an event. Moreover, our approach integrates the concept of contrastive explanation. For instance, if a symptom associated with an incorrect diagnosis has a high occurrence rate and does not appear in the clinical case, it can be invoked to discard the incorrect diagnosis. This aligns with the idea that the differences between two events form the basis for the explanation.

The computation of the prevalence function starts with the acquisition of the set of potential reasons to be evaluated. In the context of medical diagnosis, these reasons correspond to symptoms, which can be identified either within the clinical case or within the Human Phenotype Ontology (HPO) as symptoms associated with each potential diagnosis. This information facilitates the definition of three distinct sets of reasons, which serve as the basis for the computation of the Prevalence function:

- *SymptomsOfCorrectDiagnosis*: symptoms that belong to the correct disease;
- *SymptomsOfIncorrectDiagnosis*: symptoms that belong to all the incorrect diseases;
- *PresentSymptoms*: symptoms found in the case description.

The Prevalence Function is then used in conjunction with the additional disease information and symptom sets obtained from the HPO to produce a list of key reasons and their calculated prevalence scores. This allows us to provide a robust and transparent framework for assessing the quality of the reasons on which the explanations are grounded.

**Prevalence Function Algorithm** The Prevalence Score Function, as outlined in Algorithm 2, is designed to assess the relevance or pertinence of a given key reason



in the context of a clinical case (CC) and a knowledge base (KB). The function operates by assigning a score to the key reason based on its presence in the correct and incorrect diagnoses, its occurrence rate, and its presence in the clinical case.

The function begins by initializing the score to zero and setting several boolean variables to false (lines 2-6). It then retrieves the symptoms associated with the correct and incorrect diagnoses from the knowledge base (lines 7-8) and identifies the symptoms present in the clinical case using Named Entity Recognition (NER) (line 9).

The function then checks if the key reason is present in the symptoms of the correct diagnosis (lines 10-15). If it is, the function increments the score and sets the variable *linkedToCorrectDiagnosis* to true. If not, *linkedToCorrectDiagnosis* is set to false.

Next, the function checks if the key reason is present in the symptoms of the incorrect diagnoses (lines 16-20). If it is, the variable *linkedToIncorrectDiagnosis* is set to true. If not, it is set to false.

The function then checks the occurrence rate of the key reason (lines 21-28). If the key reason has a high occurrence rate (more than 70%), the variable *hasHighOccurrenceRate* is set to true, and if it is linked to the correct diagnosis, the score is incremented. If the key reason has a low occurrence rate (less than 30%), the variable *hasLowOccurrenceRate* is set to true.

The function then checks if the key reason is unique to the correct diagnosis or shared with other diagnoses (lines 29-40). If the key reason is unique to the correct diagnosis and has a low occurrence rate, the score is incremented twice. If the key reason is shared with other diagnoses, the score is decremented.

Finally, the function checks if the key reason is present in the symptoms of the incorrect diagnoses but not in the present symptoms (lines 41-48). If the key reason has a high occurrence rate, the score is incremented. Otherwise, the score is decremented.

The final score represents the prevalence of the key reason in the context of the specific clinical case, providing a measure of its relevance or pertinence.

#### 4.4.5 Reason Alignment via Sentence Matching

A crucial step in our approach is the alignment of potential causes (i.e., reasons) identified in the clinical case with those actually invoked in the expert's explanation. This alignment (visualized as the "Reasons alignment module" in Figure 4.1) is achieved through a sentence matching technique. The objective of this step is to discern which of the potential reasons identified in the clinical case were actually utilized by the experts in their explanation, thereby enabling subsequent suggestions of modifications to enhance the explanation's pertinence.

Our approach to sentence matching is inspired by the work of Lu et al. [168], particularly their creation of an intermediate dataset using a distance metric for fine-tuning their sentence-matching model. In their work, Lu et al. [168] employ

**Algorithm 3** Prevalence Function

---

```

0: procedure PREVALENCEFUNCTION(KeyReason, CC, KB)
0:   score = 0
0:   uniqueToCorrectDiagnosis = False
0:   sharedToOtherDiagnosis = False
0:   hasLowOccurrenceRate = False
0:   hasHighOccurrenceRate = False
0:   SymptomsOfCorrectDiagnosis = KB(CorrectDisease)
0:   SymptomsOfIncorrectDiagnosis = KB(IncorrectDiseases)
0:   PresentSymptoms = NER(CC)
0:   if KeyReason is in SymptomsOfCorrectDiagnosis then
0:     linkedToCorrectDiagnosis = True
0:     score = score + 1
0:   else
0:     linkedToCorrectDiagnosis = False
0:   end if
0:   if KeyReason is in SymptomsOfIncorrectDiagnosis then
0:     linkedToIncorrectDiagnosis = True
0:   else
0:     linkedToIncorrectDiagnosis = False
0:   end if
0:   if KeyReason has a high occurrence rate (more than 70%) then
0:     hasHighOccurrenceRate = True
0:     if linkedToCorrectDiagnosis == True then
0:       score = score + 1
0:     end if
0:   else if KeyReason has a low occurrence rate (less than 30%) then
0:     hasLowOccurrenceRate = True
0:   end if
0:   if KeyReason is in SymptomsOfCorrectDiagnosis then
0:     if KeyReason is not in SymptomsOfIncorrectDiagnosis then
0:       uniqueToCorrectDiagnosis = True
0:       score = score + 1
0:       if hasLowOccurrenceRate == True then
0:         score = score + 1
0:       end if
0:     else if KeyReason is in SymptomsOfIncorrectDiagnosis then
0:       sharedToOtherDiagnosis = True
0:       score = score - 1
0:     end if
0:   end if
0:   if KeyReason is in SymptomsOfIncorrectDiagnosis then
0:     if KeyReason is not in PresentSymptoms then
0:       if KeyReason has a high occurrence rate then
0:         score = score + 1
0:       else
0:         score = score - 1
0:       end if
0:     end if
0:   end if
0: end procedure=0

```

---

the Jaccard distance to identify sentences with high similarity between complex and simplified texts. We adapt this strategy to our context, aiming to locate similar reasons between the clinical case and the explanations provided by the explainer. In our adaptation of Lu et al.'s approach, we aim to identify similar reasons between the clinical case and the explanations provided by the experts. However, our methodology diverges in two key aspects: the choice of distance metric, and the preprocessing of the texts for comparison. Instead of employing the Jaccard distance, we opt for a process that begins with the detection of medical-named entities within both texts. Following this, the texts are segmented into individual sentences.

Subsequently, we compute sentence embeddings using only the identified named entities. This computation leverages the Sentence Transformers method [71], using various pre-trained models specialized in scientific text. To align sentences from the clinical case with those in the explanations, we employ cosine similarity. A match is considered valid only when the cosine distance is sufficiently close, ensuring that only highly similar sentences are matched, thereby enhancing the precision of our reason alignment process.

#### 4.4.6 Template-Based Explanation Generation

In the final step of our pipeline, we employ a template-based generation approach to articulate the pertinence of each reason. This approach allows us to generate natural language explanations that are understandable by human users. Each template is designed to address a specific combination of features associated with a reason, and the appropriate template is selected based on the values of these features for each reason. The features considered in our approach are:

- *uniqueToCorrectDiagnosis* indicates whether the reason is unique to the correct diagnosis.
- *sharedToOtherDiagnosis* indicates whether the reason is shared with other diagnoses.
- *hasLowOccurrenceRate* indicates whether the reason has a low occurrence rate.
- *hasHighOccurrenceRate* indicates whether the reason has a high occurrence rate.
- *linkedToCorrectDiagnosis* indicates whether the reason is directly linked to the correct diagnosis.
- *linkedToIncorrectDiagnosis* indicates whether the reason is linked to an incorrect diagnosis.
- *presentInClinicalCase* indicates whether the reason is present in the clinical case.

Based on the values of these features, a template, like the following, is selected to generate the explanation:

- *Template 1*: "You should consider invoking the reason [reason] since it is unique to the correct diagnosis, has a high occurrence rate, and is present in the clinical case."
- *Template 2*: "You should also consider invoking the reason [reason] since it is a symptom with a high occurrence rate for the incorrect disease [disease], and it does not appear in the clinical case, supporting discard [disease] as the correct diagnosis."
- *Template 3*: "You should consider removing the reason [reason] since it is a common symptom alongside all possible diagnoses."

This template-based generation approach allows us to generate explanations that are informative and specific to the context of each reason, thereby ensuring the interpretability of the proposed approach.

#### 4.4.7 Evaluation

In this section, we first present the experimental setting we propose to assess our approach, and then we discuss the obtained results. Finally, we apply our approach to a clinical case to discuss the final outcome of the pipeline.

**Experimental Setting.** The main experimental component of our task is the named entity-based sentence matching. This task can be decomposed into two sub-tasks: first, the generation of tuples of similar sentences based on the medical NERs, and second, the fine-tuning of Language Models (LMs) using the aforementioned dataset.

The tuple generation process begins with the segmentation of the clinical cases and the corresponding explanations into individual sentences. Subsequently, we employ the initial step of the approach proposed by Marro et al. [106], which involves the detection of medical named entities. These entities are joined into a single sentence and serve as input to the sentence embedding model [71]. We then compute cosine distances between each sentence in the clinical case and each sentence in the associated explanation. Then a dataset composed of

$$(case\_id, answer\_sentence_i, case\_sentence_j, \{True|False\})$$

tuples is generated. For the fine-tuning of the LMs, we employ the PyTorch implementation provided by Hugging Face [102]. The experiments were conducted with a batch size of 8, a maximum sequence length of 256, and a learning rate of 2.5e-5 over 4 epochs. We selected all-mpnet-base-v2 [71] as our baseline, and fine-tuned models such as BioBERT v1.2 [134], S-PubMedBert-MS-MARCO [166], and BioBERT-mnli-scinli-scitail-mednli-stsb [169] as more domain-specific LMs under the same experimental setting. Despite each transformer model achieving its best results with a

different cosine similarity threshold for performing the named entity-based matching, we kept a threshold value of 0.975 to ensure the matching of sentences with the highest possible semantic coherence.

**Results.** In this section, we present the obtained results on the Casimedicos dataset. We evaluate the quality of explanations written by experts, highlighting both successful and unsuccessful examples. The performance of the sentence matching task is quantified in terms of macro-average precision, recall, and F1 score, as shown in Table 4.5.

We adopt the all-mpnet-base-v2 model as our baseline, being it the state-of-the-art for sentence embedding computation across various domains. However, our results indicate that domain-specific models outperform this baseline across all metrics. In particular, the model based on PubMedBert [68, 166] demonstrates superior performance, achieving the highest scores in precision, recall, and F1 score (highlighted in bold in Table 4.5). These results underline the importance of domain-specific models in achieving high-quality sentence matching.

TABLE 4.5: Results for named entity-based matching in macro multi-class precision, recall, and F1-score.

Model	P	R	F1
all-mpnet-base-v2	0.84	0.70	0.74
BioBERT cased v1.2	0.84	0.76	0.80
BioBERT-mnli-snli-scinli-scitail-mednli-stsb	0.85	0.79	0.82
S-PubMedBert-MS-MARCO	<b>0.89</b>	<b>0.85</b>	<b>0.87</b>

To illustrate the outcome of our approach, we present a full clinical case, the expert’s explanation, and the assessment of reasons from the CasiMedicos dataset. We consider a clinical case where the correct diagnosis is *Porphyria cutanea tarda*. The other potential diagnoses considered are *Epidermolysis bullosa acquisita*, *Acute intermittent porphyria*, and *Ulerythema ophryogenesis*.

**Clinical Case:** “A 62-year-old man with a history of significant alcohol abuse, carrier of hepatitis C virus, treated with Ibuprofen for tendinitis of the right shoulder, goes to his dermatologist because after spending two weeks on vacation at the beach he notices the appearance of tense blisters on the dorsum of his hands. On examination, in addition to localization and slight malar hypertrichosis.”

**Expert’s Explanation:** “Porphyria Cutanea Tarda: 60% of patients with PCT are male, many of them drink alcohol in excess, women who develop it are usually treated with drugs containing estrogens. Most are males with signs of iron overload, this overload reduces the activity of the enzyme uroporphyrinogen decarboxylase, which leads to the elevation of uroporphyrins. HCV and HIV infections have been implicated in the precipitation of acquired PCT. There is a hereditary form with AD pattern. Patients with PCT present with blistering of photoexposed skin, most frequently on the dorsum of the hands and scalp. In addition to fragility, they may

develop hypertrichosis, hyperpigmentation, cicatricial alopecia, and sclerodermal induration.”

**Assessment of Reasons:** The generated explanations for the top and bottom scoring reasons are as follows:

- You should consider invoking the reason *Abnormal blistering of the skin* since it is a symptom unique to the correct diagnosis, has a high occurrence rate, and it is present in the clinical case.
- You should also consider invoking the reason *Abnormal hair morphology* since it is a symptom with a high occurrence rate for the incorrect disease *Epidermolysis bullosa acquisita*, and it does not appear in the clinical case, supporting discard *Epidermolysis bullosa acquisita* as the correct diagnosis.
- You should consider invoking the reason *Alcoholism* since it is a symptom unique to the correct diagnosis and present in the clinical case.
- The symptom *Contact dermatitis* does not meet the criteria for a strong reason in this case.
- The symptom *Dry skin* does not meet the criteria for a strong reason in this case.
- The symptom *Dermal atrophy* does not meet the criteria for a strong reason in this case.

The Expert’s Explanation for this case attributes the patient’s condition to Porphyria Cutanea Tarda (PCT), citing factors such as the patient’s gender, alcohol abuse, and the presence of blistering on photoexposed skin. These align with our top-scoring reasons in the Assessment of Reasons, demonstrating the agreement between the expert’s explanation and our assessment. In the expert’s explanation, the symptom “Abnormal hair morphology” is not mentioned. However, our methodology identifies it as a significant reason that could enhance the explanation. This symptom is common in the incorrect disease *Epidermolysis bullosa acquisita*, but it is not present in the clinical case. Therefore, its absence provides a strong reason to discard *Epidermolysis bullosa acquisita* as the correct diagnosis. This additional information could potentially enhance the expert’s explanation by providing further evidence to support the correct diagnosis and rule out other alternatives. This demonstrates the capability of our approach not only to validate the reasons used by the expert but also to suggest new pieces of information that could enrich the explanation.

## 4.5 Natural Language Explanation Generation

In the previous section, we described the first steps of our pipeline for automatically identifying the relevant symptoms which occur in the clinical case description and

then matching them with the symptoms associated with the diseases in the medical knowledge base HPO. We move now to the last step of the pipeline, i.e., the generation of natural language explanatory arguments, according to the identified relevant symptoms for the correct and incorrect diagnoses. Given the specificity of the clinical data we are dealing with, we decided to address this task by generating explanations through the definition of explanatory patterns [145–147]. We have therefore defined different patterns which take into account the different requirements of our use case scenario, where we aim at (i) explaining the correct answer by the detected symptoms and their frequency, (ii) explaining why the incorrect options cannot hold, and (iii) highlighting the relevant symptoms not explicitly mentioned in the clinical case. Let us consider the following clinical case, where in bold we highlight the **symptoms** and we underline the relevant symptoms supporting the correct answer.

**Clinical case.** A previously healthy 34-year-old woman is brought to the physician because of **fever** and **headache** for 1 week. She has not been exposed to any disease. She takes no medications. Her temperature is 39.3°C (102.8°F), pulse is 104/min, respirations are 24/min, and blood pressure is 135/88 mm Hg. She is **confused** and oriented only to person. Examination shows jaundice of the skin and **conjunctivae**. There are a few scattered **petechiae** over the trunk and back. There is **no lymphadenopathy**. Physical and neurologic examinations show **no other abnormalities**. Test of the stool for occult blood is positive. Laboratory studies show: Hematocrit 32% with fragmented and nucleated erythrocytes Leukocyte count 12,500/mm<sup>3</sup> Platelet count 20,000/mm<sup>3</sup> Prothrombin time 10 sec Partial thromboplastin time 30 sec Fibrin split products negative Serum Urea nitrogen 35 mg/dL Creatinine 3.0 mg/dL Bilirubin Total 3.0 mg/dL Direct 0.5 mg/dL Lactate dehydrogenase 1000 U/L Blood and urine cultures are negative. A CT scan of the head shows **no abnormalities**. Which of the following is the most likely diagnosis?

The correct diagnosis is Thrombotic thrombocytopenic purpura, whilst the other (incorrect) options are Disseminated intravascular coagulation, Immune thrombocytopenic purpura, Meningococcal meningitis, Sarcoidosis and Systemic lupus erythematosus.

**Why Pattern.** We focus here on the correct diagnosis explanation pattern, which allows explaining why this is the correct diagnosis. We define the following template to generate our natural language explanations:

**Definition 1** (*Why for correct diagnosis*) The patient is showing a [CORRECT DIAGNOSIS] as these following symptoms [**PERFECT MATCHED SYMPTOMS**, **MATCHED SYMPTOMS**] are direct symptoms of [CORRECT DIAGNOSIS].

Moreover, [**OBLIGATORY SYMPTOMS**] are obligatory symptoms (always present, i.e., in 100% of the cases) and [**VERY FREQUENT SYMPTOMS**] are very frequent symptoms (holding on 80% to 99% of the cases) for [CORRECT DIAGNOSIS] and are present in the case description.<sup>6</sup>

In Template 1, the [CORRECT DIAGNOSIS] represents the correct answer to the question "Which of the following is the most likely diagnosis?" and therefore the

<sup>6</sup>Sources from HPO: <https://hpo.jax.org/app/browse/term/HP:0040279>

correct diagnosis of the described disease. The **[SYMPTOMS]** in bold represent the symptoms automatically detected through the first module of our pipeline, and they are also underlined when they are considered as relevant by our matching module, i.e., they are listed among the symptoms for the disease in the HPO knowledge base. Both **[PERFECT MATCHED SYMPTOMS]** and **[MATCHED SYMPTOMS]** in Template 1 are considered relevant but they differ in the confidence level the system assigns to the matched symptoms. This allows us to integrate a notion of granularity in our explanations and to rely on the symptoms detected in the clinical case that strongly match with a symptom in HPO. If the system does not detect any relevant symptom, no explanation is generated for the correct answer. Furthermore, we employ the information about the symptom frequencies (retrieved through HPO) in the **[OBLIGATORY SYMPTOMS]** and **[VERY FREQUENT SYMPTOMS]** to generate stronger evidence to support our natural language argumentative explanations. Sometimes the frequencies are not available in the HPO, in which case we do not display them in our final explanation.

We present now some examples of explanatory arguments automatically generated by our system.

**Example 4.5.1** *The patient is showing a [Thrombotic thrombocytopenic purpura] as these following symptoms [Headache, Fever, Confusion (Oriented to persons) and Reticulocytosis (Jaundice of the skin)] are direct symptoms of [Thrombotic thrombocytopenic purpura].*

*Moreover [Reticulocytosis (Jaundice of the skin)] are very frequent symptoms (holding on 80% to 99% of the cases) for [Thrombotic thrombocytopenic purpura] and are present in the case description.*

When filling the **[SYMPTOMS]** span in Template 1, we inject only the symptoms matched in the HPO for the **[PERFECT MATCHED SYMPTOMS]**, and we combine the HPO symptoms with the symptoms detected in the case description for the **[MATCHED SYMPTOMS]** in this form: **[matched symptom in HPO (detected symptom in the clinical case)]** (e.g., in Example 4.5.1:

**Confusion (Oriented to persons) and Reticulocytosis (Jaundice of the skin)**)

**Why not Template.** Explaining why a diagnosis is the correct one is important, but it is also necessary to be able to say why the other options are not correct as possible diagnoses for the clinical case under investigation [108]. We, therefore, propose to provide explanations based on the relevant symptoms for the incorrect options by contrasting them with the clinical case at hand.

**Definition 2** (*Why not for incorrect diagnosis*) *Concerning the [INCORRECT DIAGNOSIS] diagnosis, it has to be discarded because the patient in the case description is not showing [INCORRECT DIAGNOSIS SYMPTOMS FROM HPO (MINUS DETECTED SYMPTOMS IN CASE)] symptoms.*

*Despite [SHARED CORRECT SYMPTOMS] symptoms shared with the [CORRECT DIAGNOSIS] correct diagnosis, the [INCORRECT DIAGNOSIS] diagnosis is based on [INCORRECT DIAGNOSIS SYMPTOMS].*



Moreover, [**OBLIGATORY SYMPTOMS**] are obligatory symptoms (always present, i.e., in 100% of the cases) and [**VERY FREQUENT SYMPTOMS**] are very frequent symptoms (holding on 80% to 99% of the cases) for [**INCORRECT DIAGNOSIS**], and they are not present in the case description.

Template 2 can be applied to each incorrect possible answer of the case, individually. The incorrect answer corresponds to the [**INCORRECT DIAGNOSIS**] and [**INCORRECT DIAGNOSIS SYMPTOMS**] are all relevant symptoms associated with this disease in the HPO knowledge base, without the symptoms in common with the correct answer. Again, in the template, we use the frequencies provided by HPO to provide further evidence to make our explanatory arguments more effective. The template includes therefore with [**OBLIGATORY SYMPTOMS**] and [**VERY FREQUENT SYMPTOMS**] the mandatory and very frequent symptoms of the incorrect diagnosis, which are missing in the clinical case description. The following explanations are automatically generated for (one of) the incorrect diagnoses of the clinical case we introduced at the beginning of this section.

**Example 4.5.2** Concerning the [*Meningococcal meningitis*] diagnostic, it has to be discarded because the patient in the case description is not showing [*Stiff neck, Nuchal rigidity or CSF pleocytosis, Increased CSF protein, Hypoglycorrhachia*] symptoms.

Despite [*Petechiae, Fever, Headache*] symptoms shared with the [*Thrombotic thrombocytopenic purpura*] correct diagnosis, the [*Meningococcal meningitis*] diagnosis is based on [*Stiff neck, Nuchal rigidity or CSF pleocytosis, Increased CSF protein and Hypoglycorrhachia*].

Moreover, [*Stiff neck, Nuchal rigidity, CSF pleocytosis, Increased CSF protein or Hypoglycorrhachia*] are very frequent symptoms (holding on 80% to 99% of the cases) for [*Meningococcal meningitis*] and are not present in the case description.

Example 4.5.2 shows the NL explanation of why the possible answer [*Meningococcal meningitis*] is not the correct diagnosis given the symptoms discussed in the clinical case description. In case the disease is not found in HPO, we do not generate the associated explanation.

**Additional Explanatory Arguments.** In order to enrich our explanations with additional explanatory arguments to improve critical thinking in the medical residents, we also generate another template. Indeed, in some clinical cases, it is possible that the symptoms are not sufficient to explain the diagnosis or sometimes the symptom has to be combined with vital signs or other characteristics of the patient to be correctly interpreted. Some of these signs represent potentially important symptoms for the diagnosis, as in the previous example, where the sentence *respirations are 24/min* could be associated with the symptom of *Dyspnea* in HPO. Template 3 aims at drawing the medical residents' attention to (statistically) important symptoms that are missing or not explicitly mentioned in the clinical case description:

**Definition 3** Furthermore, [**CORRECT DIAGNOSIS VERY FREQUENT SYMPTOMS (MINUS MATCHED SYMPTOMS)**] are also frequent symptoms for [**CORRECT DIAGNOSIS**] and could be found in the findings of the clinical case.

Example 4.5.3 is generated by our system and brings attention to Dyspnea. This additional explanatory argument complements the explanation we generate for the correct and incorrect diagnoses in the case presented at the beginning of this section.

**Example 4.5.3** *Furthermore, [Dyspnea, Thrombocytopenia, Generalized muscle weakness, Reticulocytosis, and Microangiopathic hemolytic anemia] are also frequent symptoms for [Thrombotic thrombocytopenic purpura] and could be found in the findings of the clinical case.*

## 4.6 Concluding Remarks

In this chapter, a pipeline is presented to generate natural language explanatory arguments for correct and incorrect diagnoses in clinical cases. The pipeline first automatically identifies relevant symptoms in a clinical case description and matches them to medical knowledge base terms to associate symptoms with potential diagnoses. It then generates explanatory arguments highlighting *why* one diagnosis is deemed correct and others incorrect. Experiments on a dataset of 314 clinical cases in English demonstrate promising results, with 0.86 F1-Score for symptom detection and 0.53 accuracy for top-5 symptom matching, outperforming competitive baselines and state-of-the-art approaches.

Furthermore, an approach is presented to recognize the cognitive processes underlying explanation selection and aim to emulate them transparently. By incorporating principles of abnormality and contrastive explanation, the approach is attuned to real-world explanation selection, with a focus on medicine. By leveraging the Human Phenotype Ontology and named entity recognition, potential reasons are identified and assessed systematically to evaluate explanation alignment with expert perspectives and suggest enhancements.

However, limitations exist. Template-based explanation generation has drawbacks like design dependence and inflexibility that may reduce effectiveness in dynamic settings. Nevertheless, as symptoms are stable data, this is less concerning in the current medical application. Future work could investigate conversational systems and adaptive explanation strategies customized to users' knowledge. Expanding the approach to multiple knowledge bases could incorporate more external evidence to further validate claims. Overall, the proposed techniques demonstrate promising capability in extracting, aligning, and assessing clinical arguments to apply external medical knowledge. Opportunities remain to enhance reasoning transparency through expanded knowledge and personalized interaction.



## Chapter 5

# Argument Mining on Clinical Trials

*This chapter introduces the different applications I contributed to during my Ph.D. These applications span from the Covid-on-the-Web project developed to address an argumentative analysis of the clinical articles about COVID-19 to the study of the effect of interventions on the outcomes for the AbsRCT dataset [22] towards the implementation of a web tool that provides argumentative analysis of medical articles to support clinicians decision making in real-time. This chapter surveys on the results published at the International Semantic Web Conference (ISWC-2020) [170], in Artificial Intelligence in Medicine journal (Elsevier 2021) [22] and in the Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2022) [171] (demo paper).*

This chapter details contributions applying argument mining techniques to real-world medical scenarios. First, an overview is provided of ACTA [172], an automated tool supporting clinicians in extracting argument graphs from clinical trial reports. ACTA was then utilized in the Covid-on-the-Web project to transform unstructured coronavirus literature into rhetorical structures, aiding evidence-based reasoning. To enhance ACTA's capabilities, the AbsRCT dataset [22] was extended with annotations on medical intervention outcomes and their effects on patients. Methods were developed to automatically classify nuanced effect types like Increased or Decreased. This argument mining tool extracts structured results data from trials to help clinicians efficiently interpret findings. Subsequently, ACTA underwent upgrades to version 2.0, integrating state-of-the-art neural models, a new module to analyze outcome effects, and a modular API-based architecture for customization. Overall, ACTA 2.0 pushes boundaries in clinical argument mining to empower evidence-based medicine through automated search, extraction, and visualization. The chapter discusses each application in turn, delineating how argument mining techniques were specialized and applied to advance real-world medical scenarios like infectious disease research, drug trials, and decision support. By targeting healthcare's intricacies, the work reveals practical benefits across domains

while tackling open technical challenges in knowledge-aware language understanding, reasoning modelling, and adaptable engineering.

In this context, Mayer et al. [172] presented ACTA, a tool to help clinicians analyze clinical trial arguments by extracting claims, evidence, and relations. The overarching goal is to transform unstructured textual trial reports into structured argument graphs. This provides a summary of key claims, evidence, and relations to assist evidence-based decision-making. The ACTA tool is designed to search for trial abstracts on PubMed. Once a user selects an abstract, the text is processed using natural language processing techniques, specifically, argument component detection. This involves identifying claims and premises, which are then marked as argument span boundaries using BIO notation. To handle this tagging, a pre-trained BERT model is fine-tuned. To predict relations, a choice classification approach is used. The model selects the most likely target component that a source component connects to, from a list of candidates. This ensures that there is at most one outgoing edge per node when constructing the final graph. Key outputs include highlighting detected arguments and PICO elements in the original text, and visualizing the argument graph. However, ACTA faced limitations like handling clinical abbreviations and lacking more nuanced relation types beyond binary link prediction. ACTA pioneered automated analysis of clinical trial arguments to support evidence-based medicine. In Section 5.1 I detail how ACTA was utilized in the Covid-on-the-web project to extract argumentative structures from coronavirus literature. However, opportunities remained to upgrade the techniques and expand the functionality. Subsequent sections will detail ACTA 2.0.

## 5.1 Covid-on-the-Web project

In Spring 2020, the rapid spread of the novel coronavirus SARS-CoV-2 motivated my research team at Inria<sup>1</sup> to join global efforts to fight the pandemic. We launched the *Covid-on-the-Web* project to assist in utilizing scientific literature on coronaviruses. During the initial pandemic lockdowns, we adapted and combined our existing methods, models, and tools (ACTA, Corese<sup>2</sup> [173], MGExplorer [174], Morph-xR2RMLL<sup>3</sup>) to process, analyze, and enrich the "Covid-19 Open Research Dataset" (CORD-19), which contains over 50,000 coronavirus-related scientific articles.

The overarching goal of Covid-on-the-Web is to make Covid-19 literature more accessible and useful for biomedical researchers. We designed a pipeline to continuously enrich a knowledge graph on COVID-19, as well as software to exploit this graph. Our approach leverages knowledge representation, text/data/argument mining, data visualization, and exploration. The pipeline extracts named entities mentioned in the articles, drawing from DBpedia, Wikidata, and other BioPortal

<sup>1</sup><https://team.inria.fr/wimmics/>

<sup>2</sup><https://project.inria.fr/corese/>

<sup>3</sup><https://github.com/frmichel/morph-xr2rml/>

vocabularies. It also extracts argumentative graphs to help clinicians analyze clinical trials and make evidence-based decisions. On top of the knowledge graph, we developed and deployed tools for visualization, exploration, and data science notebooks.

We engaged with biomedical institutions, including project partners like the French National Institute of Health and Medical Research (Inserm)<sup>4</sup> and the French National Cancer Institute (INCa), to ensure our approach aligned with real needs. Through discussions, we identified motivating scenarios and competency questions to guide and test our knowledge graph. For example, users suggested queries like "Find all articles discussing both a cancer type and a corona-type virus." Continued elicitation of meaningful queries helps specify and validate our resources.

The Covid-on-the-Web pipeline and services aim to address key scenarios:

- Helping clinicians analyze clinical trials and make evidence-based decisions using argumentative graphs.
- Assisting hospital physicians in collecting normal ranges for biomarkers from scientific articles.
- Enabling cancer institute researchers to find articles on cancer-coronavirus links to inform research programs.

A key pipeline component involves creating RDF argumentative subgraphs with the ACTA [172] tool. As introduced before in Section 2.1 ACTA identifies the different argumentative components such as claims or premises, among their corresponding support or attack relationships. For example, ACTA could extract an argument with the claim "Hydroxychloroquine is an effective COVID-19 treatment" supported by a small trial's results. Clinicians can then critically assess the argument's cogency. By transforming unstructured coronavirus literature into meaningful argument graphs, the Covid-on-the-Web pipeline aims to make evidence-based reasoning more efficient.

In this section, I describe the pipeline developed through the Covid-on-the-Web project to generate Linked Data from the CORD-19 dataset. I first provide a high-level overview of the pipeline architecture and its various components and functionalities. I then delve into the details of how the pipeline creates argumentative RDF subgraphs using the ACTA tool.

### 5.1.1 The Covid-on-the-Web RDF Dataset

To create meaningful Linked Data about coronavirus we analyzed the unstructured COVID-19 Open Research Dataset (CORD-19) [175]. Applying ACTA to enrich this dataset with argumentative graphs was one of the many steps we took in order to create the RDF. In this context, the expertise of the team with respect to knowledge graphs allowed us to enrich the CORD-19 dataset from many different sources.

---

<sup>4</sup><https://www.inserm.fr/>

Named entity recognition tools like DBpedia Spotlight [176], Entity-fishing<sup>5</sup> and NCBO BioPortal Annotator [177] were employed to identify biomedical entities in the articles. Such tools allowed us to disambiguate them against Linked Open Data (LOD) resources from DBpedia, Wikidata and BioPortal ontologies.

The output of these NLP tools, along with the argument graphs from ACTA, are converted to RDF triples using the Morph-xR2RML framework<sup>6</sup>. Such framework maps the extracted information into an RDF representation using ontologies and vocabularies that make the data interoperable on the Semantic Web. For example, an extracted claim like "Hydroxychloroquine treats COVID-19" would be encoded with triples asserting the "treats" relationship between the two named entities.

The resulting knowledge graph, called the *Covid-on-the-Web RDF dataset*, is served through a public SPARQL endpoint allowing structured queries. We paid particular attention to open and reproducible science principles, making the data, code, and process fully transparent and reusable. The RDF representation enables interoperability with other coronavirus datasets. However, a unique capability is the inclusion of argumentative structures from ACTA alongside biomedical entities.

To manipulate this knowledge graph, we integrated visualization platforms like Corese<sup>7</sup> [173] and MGExplorer [174]. These tools help users analyze relationships in query results, aiding understanding. MGExplorer and the enclosed notebooks also enable data scientists to transform results into analysis-ready structures. The customizable visualization widgets allow institutions to tailor interfaces to their own scenarios and competency questions. As shown in Figure 5.1, all of these tools were fused into a single pipeline [170]. The modular nature of the tools integrated into the Covid-on-the-Web pipeline allows for the flexible application of its linked data resources to diverse biomedical scenarios beyond the current focus on SARS-CoV-2 literature. The interoperability provided by representing extracted information as structured RDF graphs enables extensions of the knowledge graph to incorporate new knowledge sources. In this way, the generality of our approach means the pipeline's utility may continue even as the focus shifts to future challenges facing researchers and clinicians.

As new COVID-19 versions are released, the RDF dataset is continuously updated to stay current. This supports monitoring how biomedical knowledge evolves throughout the pandemic. Overall, the semantically structured knowledge graph enhances access to the exponentially growing corpus of coronavirus literature. Linking rhetorical structures and biomedical entities also facilitates evidence-based reasoning and aids clinicians in making sense of this complex, unfinished science.

<sup>5</sup><https://github.com/kermitt2/entity-fishing>

<sup>6</sup><https://github.com/frmichel/morph-xr2rml/>

<sup>7</sup><https://project.inria.fr/corese/>

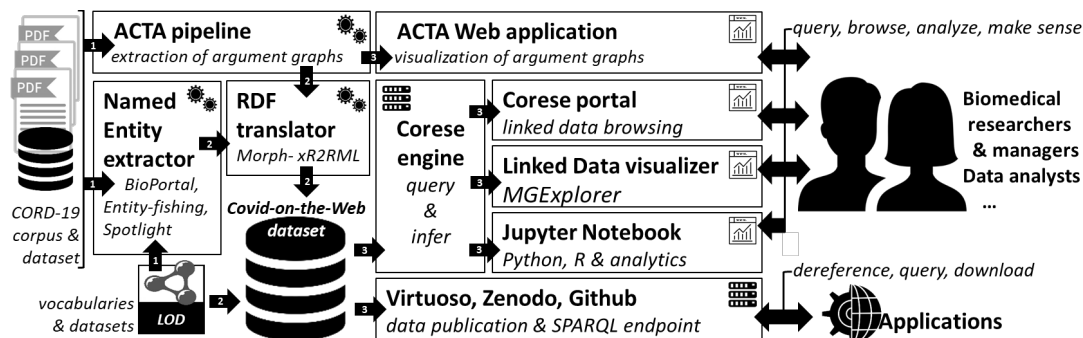


FIGURE 5.1: Illustration of the Covid-on-the-Web [170] pipeline, its services and applications.

### CORD-19 Argumentative Knowledge Graph

The Covid-on-the-Web pipeline generates two interconnected knowledge graphs from the CORD-19 dataset: the Named Entities Graph and the Argumentative Graph. This section focuses on the creation of the Argumentative Graph using the ACTA tool. In brief, ACTA analyzes each CORD-19 abstract to extract argumentative structures - claims, evidence, and their relationships. It then represents these structures in RDF format using established argumentation ontologies. The pipeline involves four main steps:

**Component Detection** As detailed in Section 2.1, ACTA views this as a sequence tagging task, using a transformer model to identify argument components in the text. In this scenario, pre-trained SciBERT weights were used instead of BERT in order to improve performance given the scientific language.

**Relation Classification** The type of relationship between argument components is determined via a 3-class sequence classification approach. A fine-tuned SciBERT transformer model creates numerical representations of the input text, which consists of component pairs. A linear classification layer then predicts if the relationship is support, attack or no relation. This extends the original ACTA pipeline by moving beyond binary link prediction to a richer, nuanced graph showing important distinctions like evidence challenging or backing a claim.

**PICO Extraction** Using the same model as in ACTA, PICO elements are detected within argument components only, not whole abstracts. Unique concepts are then linked to UMLS for standardized representation.

**RDF Conversion** Argumentative information is structured in RDF format using the Argument Model Ontology (AMO)<sup>8</sup>, the SIOC Argumentation Module (SIOCA)<sup>9</sup>

<sup>8</sup><http://purl.org/spar/amo/>

<sup>9</sup><http://rdfs.org/sioc/argument#>



and the Argument Interchange Format<sup>10</sup> ontologies. This enables interoperability.

Overall, ACTA automatically extracts argument graphs from unstructured CORD-19 text and represents them as meaningful, shareable RDF knowledge graphs. This facilitates evidence-based reasoning and analysis by biomedical experts.

## 5.2 Argument Mining on Clinical Trials: Computing the Effect-on-Outcome

One of the argument mining applications developed was the automated analysis of the effect of medical interventions on outcomes, termed *Effect-on-Outcome*. In clinical trial reports, it is crucial to understand how a new treatment impacts disease symptoms and side effects compared to conventional therapy. However, manually analyzing these effects is tedious and time-consuming given the volume of literature. Automating this process would provide structured results data to help clinicians efficiently interpret trials and make evidence-based decisions. This section will describe the Effect-on-Outcome work, outlining the key tasks, data, and results.

For example, consider the sentence: "Patient-reported nausea decreased after taking drug A compared to placebo, while liver enzyme levels increased." Here, the outcomes are "nausea" and "liver enzyme levels," with effects "decreased" and "increased" respectively. Automatically extracting such effects allows clinicians to quickly synthesize the trial outcomes instead of reading lengthy reports.

To enable automated outcome analysis, a two-part argument mining pipeline was developed:

**Outcome extraction** : This first stage treats outcome detection as a sequence tagging problem using a BIO notation, resulting in a three-class classification problem (*B-Outcome*, *I-Outcome* and *NoOutcome*). Different pre-trained transformer-based models adapted for sequence tagging are then finetuned, labelling the outcome spans in text, e.g. tagging "nausea" as B-Outcome and "liver enzyme levels" as B-Outcome I-Outcome I-Outcome.

**Effect classification** : The extracted outcomes are then paired with the component it occurred in, and fed into an effect classifier to predict the effect. Sentences with multiple detected outcomes generated multiple inputs, one for each detected outcome. Using labels *Improved*, *Increased*, *Decreased*, *NoDifference*, and *NoOccurrence*, several transformer-based models fine-tuned on these five classes determine the effect on each outcome.

Figure 5.2 illustrates the role of the outcome analysis in the overall AM pipeline.

The five effect classes represent different types of impacts an intervention can have on an outcome. The *Improved* class is used when the outcome had a beneficial

<sup>10</sup><http://www.arg.dundee.ac.uk/aif#>

Class	#outcomes	%
Improved	831	25
Increased	765	23
Decreased	782	23
NoDifference	897	27
NoOccurrence	76	2

TABLE 5.1: Statistics of the Outcome dataset, showing the numbers of *Improved*, *Increased*, *Decreased*, *NoDifference* and *NoOccurrence* classes independent of the disease-based subsets.

effect but the direction is unclear. *Increased* and *Decreased* indicate the outcome measure went up or down, respectively. *NoDifference* denotes no change in the outcome or no difference between arms. Finally, *NoOccurrence* is used when the outcome did not occur, typically for adverse events. By predicting these nuanced labels, the model can capture the effect an intervention had on key outcomes in a structured way.

To evaluate the Effects on Outcome task, we focus on the sentences of the AbstrCT dataset containing outcomes (i.e., 3351 sentences annotated with five classes, as reported in Table 5.1).

### 5.2.1 Experimental Setup

Experiments are conducted with the pre-trained transformer models **BERT**<sub>base</sub>, **BioBERT** and **SciBERT** (cased and uncased).

For both parts of the pipeline, i.e., the outcome detection and effect classifier, the same type of transformer is employed. As for the sequence tagging architecture the LSTM combination with a CRF was chosen for the experiments, because the difference between the LSTM and GRU approaches were only marginal for the argument component detection. The outcome pipeline implementation was done

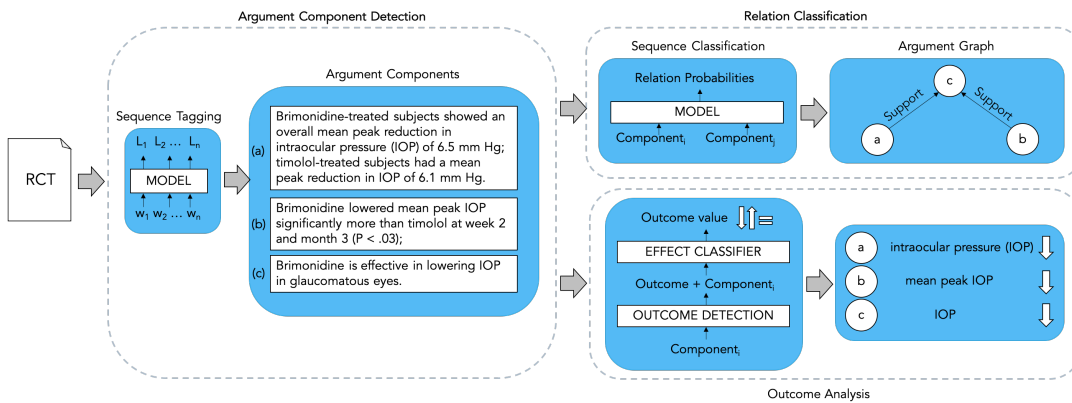


FIGURE 5.2: Illustration of the full Argument Mining pipeline with the outcome analysis extension.

with the use of the PyTorch implementation of huggingface Transformers library<sup>11</sup> version 2.3. Both transformer models of the pipeline are of the same type and initialized with the same pre-trained weights. The Effect-on-Outcome annotations are converted into two datasets, one for each part of the pipeline. The first one is in a CoNLL format for token-wise labels, and the second one is in csv format, where each outcome-component pair is listed. This results in multiple entries, if a component contains more than one outcome. The fine-tuning of the models is done separately, each task on its own dataset version. The learning rate was set to  $2e-5$  with Adam optimizer and the models were fine-tuned over 3 epochs with a batch size of 32 and a maximal sequence length of 128 tokens. Token-wise evaluation is done on the full pipeline output, which is reconverted to CoNLL format to compare against the gold labels, taking the propagated error from the first pipeline part into account. The annotated dataset was split into a train and test set (80% and 20%, respectively) respecting the class distribution of the overall dataset in both subsets.

## 5.2.2 Results and Discussion

Model	$F_1$	Improved	Increased	Decreased	NoDiff	NoOcc
BERT (cased)	.62	.69	.65	.66	.75	.00
BERT (uncased)	.72	.72	.70	.72	.72	.50
BioBERT	.75	.74	.74	.77	.76	.54
SciBERT (cased)	.75	.71	.71	.73	.71	<b>.65</b>
SciBERT (uncased)	<b>.80</b>	<b>.81</b>	<b>.75</b>	<b>.81</b>	<b>.85</b>	.59

TABLE 5.2: Results for the outcome analysis pipeline, given in overall macro  $F_1$  and label-wise binary  $F_1$ -score.

Results on the full pipeline can be seen in Table 5.2. We can observe an increase in performance on the specialized Bio- and SciBERT models compared to the general BERT model. In a direct comparison of the cased versions of these two specialised models, the overall  $F_1$ -score is the same with .75. In the binary evaluation, BioBERT is slightly better with the exception of the *noOccurrence* class. Interestingly here, the SciBERT cased model performs the best with an  $F_1$ -score of .65. Overall, SciBERT uncased is the best-performing model with a macro  $F_1$ -score of .80. It also outperforms the rest of the approaches in every  $F_1$ -score measured except for the *noOccurrence* category, where the cased version has the higher score. This category, in particular, suffers from sensitivity to class imbalance given that only 2% of the annotated data is labelled as such. For the other classes, the binary  $F_1$ -scores are in a comparable range to each other, where the most prominent class in the annotated data, i.e., *noDifference* with 27%, has consistently the highest or second highest score. Besides the *noOccurrence* class, the *Increased* class has always the second lowest scores. Even for the best-performing model, the difference compared to the worse-performing models is not as massive as for the other classes. Notable in the confusion matrix, visualized

<sup>11</sup><https://github.com/huggingface/transformers>

in Figure 5.3, the classifier tends to wrongly predict it as *Improved*, which is a closely related class. The F1-score for the overall performance of the pipeline, i.e., with the argument component detection as a prior step, is .62 for the 50% and the 100% threshold. Both constraints produce a similar F1-score. Taking a look at the number of detected components for each of the constraints, there is only a total difference of 2 between them. Varying the threshold does not change the difference by much. We found that if the model detects a component most of the time at least 70% of the tokens are detected. Concerning the strong decrease from the gold label to the overall pipeline performance, we found that the *NoOccurrence* is the main reason, with not a single sample correctly predicted; either through not finding the component or, if detected, misclassifying the outcome with the wrong label. A similar situation was observed for the BERT based model on the gold standard, where the 0 F1-score of the *NoOccurrence* class lowered the macro F1-score significantly with respect to the other models. Ignoring the *NoOccurrence* class to estimate a performance value for the other classes, the macro F1-score would be at .74 for the whole pipeline.

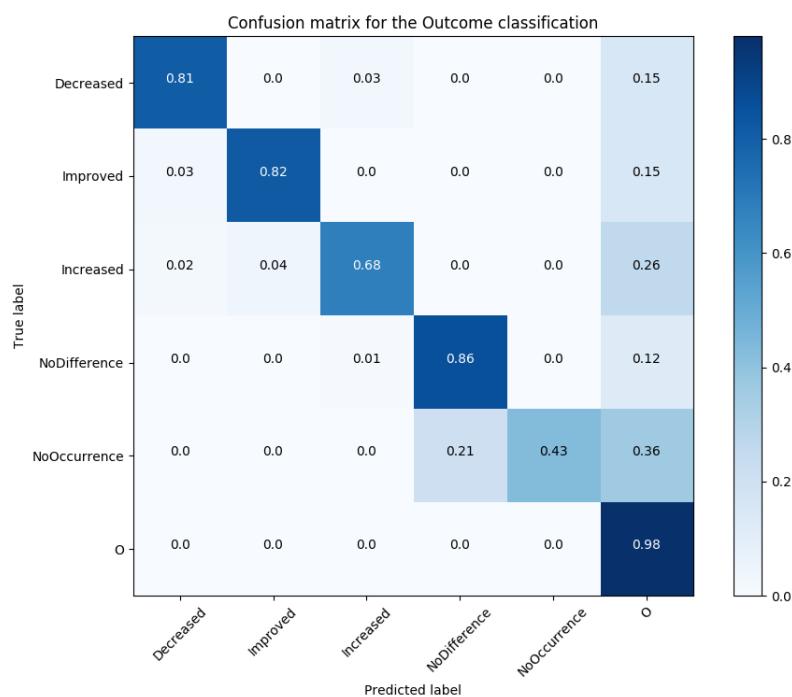


FIGURE 5.3: Confusion matrix of the predictions on the test set of the outcome classification.

**Error Analysis** With respect to the source of error in the pipeline, the two pipeline parts cause different observable errors in the overall output. Being a binary classifier, the outcome detection is the only part which predicts the negative class label (referred to as *O* in the confusion matrix). The second part, the effect classifier, assigns effect class labels (*Increased/Decreased*, etc.) to outcomes, which were found by the outcome detection module. Consequently, the impact of the propagated error

from the first part of the pipeline can be observed in the confusion matrix in Figure 5.3. Effect classes are mostly not misclassified as other effect classes, but as the negative class *O*. This is reflected in a stronger coloration in the horizontal direction for the predicted *O* label in the confusion matrix. Since the only part in the pipeline which is responsible for the negative *O* label is the outcome detection, this means that the error occurred in the first part of the pipeline. Accordingly, confusion of effect class labels are errors in the second part, the effect classifier, in the pipeline.

One of the most common mistakes of the models is the incomplete detection of outcomes. In many cases, the outcome to classify includes other words that complement it, for example in the sentence *The levels of VEGF were significantly lower*, the outcome to classify is *The levels of VEGF* while the model only catches *VEGF*. We also find that the model is effectively tagging outcomes in such a way that is different from the true labels, but correct nonetheless. For example, consider the sentence *Excess limb size (circumference and water displacement) and excess water composition were reduced significantly*. This sentence has as true labels the outcomes *Excess limb size* and *excess water composition*, both labeled as *Decreased*. The model detects and classifies those outcomes correctly, but also adding the words *circumferences* and *water displacement*, predicting the label *Decreased* which would be the correct label.

### 5.3 ACTA 2.0

Building automated tools to assist clinicians in analyzing medical literature poses unique challenges. Clinical texts require specialized language processing to handle domain terminology and parsing the rhetorical structure demands accurate extraction of claims, evidence, and relations forming argument graphs. Moreover, a usable interface must integrate search functionality while visualizing extracted semantics. Despite these difficulties, such tools offer immense potential to enhance evidence-based reasoning and decision-making.

To address these challenges, I collaborated to develop ACTA 2.0 [12], an upgraded system for argument mining in clinical trials integrating recent advances in argument mining and outcome analysis. The overarching motivation was assisting clinicians in extracting key claims and evidence from trial reports through automated argument analysis. This required tackling issues like clinical abbreviations and integrating up-to-date neural models to enhance extraction accuracy. Additionally, ACTA 2.0 aimed to provide richer contextual information by detecting PICO elements, labelling relation types and incorporating the effect-on-outcome analysis presented earlier in this thesis as a new module in the argument mining pipeline. This allows enriching extracted arguments with insights into how interventions impact disease symptoms or side effects. On the implementation side, adopting a modular API-based architecture improved flexibility and customization. Overall,

---

<sup>12</sup><http://ns.inria.fr/acta/>

ACTA 2.0 pushes boundaries in clinical argument mining to empower evidence-based medicine through automated search, extraction, and visualization. The subsequent sections detail the components enabling these key functionalities.

To the best of our knowledge, ACTA 2.0 is the only automated tool which allows for a deep analysis of clinical text from the argumentative point of view to support evidence-based medicine. Few systems tackle similar tasks, like EVIDENCEMINER [178] (which, given a natural language query, automatically retrieves sentence-level textual evidence from a corpora of biomedical literature), RobotReviewer [179, 180] (which summarizes the key information of a clinical trial, including the interventions, trial participants and risk of bias), and ExaCT [181] (which extracts information containing PICO elements based on a SVM). Also, Lehman et al. [182] proposed an approach to infer if a study provides evidence with respect to a given intervention, comparison intervention and outcome. However, none of these systems is able to extract a full argument graph (where evidence and claims are the nodes, and attacks and supports are the labelled edges) from a clinical text. Concerning the identification of PICO elements in text, different approaches are proposed in the literature [183–185] to identify them in text, but none of these approaches tackles the issue of analysing the effects of an intervention on the outcomes of a clinical trial study, as in ACTA 2.0.

### 5.3.1 Main Functionalities

ACTA 2.0 provides the following functionalities:

**Search on Pubmed.** PubMed<sup>13</sup> is a free search engine accessing primarily the MEDLINE database<sup>14</sup> of references and abstracts on life sciences and biomedical topics. Given the importance of this search engine in the healthcare domain, ACTA 2.0 maintains the possibility to search for a (set of) abstract(s) directly on the PubMed catalogue, through their API<sup>15</sup>. As in the previous version of ACTA, this API is integrated as a search bar to enter queries in the common PubMed format, similar to the original PubMed web interface. After the query is executed, when the results are shown, the user can then select one or more abstracts to proceed with the analyses offered by ACTA 2.0. Alternatively, the system accepts raw text as input to be processed via *Analyse Custom Text*.

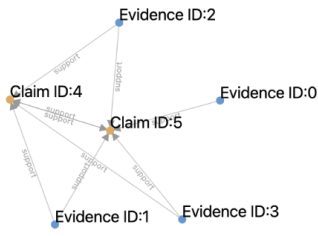
**Enhanced Argumentative Analysis.** Once the text is uploaded or an abstract is selected from the list of search results, the user can proceed with the argumentative and outcome analyses by pressing the *Analyse* button. After a short processing time, the result is visualized in the user interface in form of an argumentative graph.

<sup>13</sup><https://pubmed.ncbi.nlm.nih.gov/>

<sup>14</sup><https://www.nlm.nih.gov/medline/medlineoverview.html>

<sup>15</sup><https://pubmed.ncbi.nlm.nih.gov/advanced/>

### Argument Graph Download



**PMID** 25784905

**Title:** VgrG2 of type VI secretion system 2 of *Vibrio parahaemolyticus* induces autophagy in macrophages.

**Authors:** Yu Y, Fang L, Zhang Y, Sheng H, Fang W

**Abstract:** Type VI secretion system (T6SS) is a macromolecular transenvelope machine encoded within the genomes of several proteobacteria species. *Vibrio parahaemolyticus* contains two putative T6SS systems, VpT6SS1 and VpT6SS2, both contributing to adherence to Caco-2 and/or HeLa cells. However, it remains unknown if these systems are involved in cellular responses. In order to exclude the effects of other virulence factors known to induce cytotoxicity or autophagy, a triple deletion mutant dttt (with deletion of *tdh*, and *t3ss1* and *t3ss2* structural protein genes) was used as the parent strain to construct deletion mutants of *t6ss* genes. The mutant dttt - *δicmf2*, but not dttt - *δicmf1*, reduced autophagic response upon 4 h of infection of the macrophage. Further attempt was made to search for the possible effector proteins that might be responsible for direct induction of autophagy by deletion of the genes encoding *Hcp2* and *VgrG2*, two putative translocators of T6SS2 of *V. parahaemolyticus*. Deletion of either *hcp2* or *vgrg2* did reduce the autophagic response. However, increased *lc3 - ii lipidation* was seen only in the macrophage cells transfected with *pvgrg2*, but not with *phcp2*. Chloroquine treatment increased accumulation of *lc3 - ii*, suggesting that *vgrg2* enhanced autophagic flux. The fact that *vgrg2* deletion led to reduced level of intracellular camp suggests a possible role of camp signaling in autophagic responses to the bacterium. We conclude that VgrG2 of *V. parahaemolyticus* induces autophagy in macrophages.

**Colors code:** increased Decreased Improved No difference No occurrence

**Highlight** Argumentative Components PICO Elements Effects on Outcomes Reset text

### Effect on Outcomes

Effect	Outcome
NoDifference	cytotoxicity
Decreased	autophagic response
Decreased	autophagic response
Increased	lc3 - ii lipidation
Increased	accumulation of lc3 - ii
Increased	autophagic flux
Decreased	intracellular camp

### ACTA Argument mining tool 1.0.0 OAS3

This is the demo API of the argument mining tool ACTA <http://ns.inria.fr/acta/>. For this sample you can send your text to the server and get the arguments and links between.

[Contact the developer](#)  
Apache 2.0

**Servers**

**argument** Returns arguments and relations Find out more: <https://ns.inria.fr/acta/doc/> ^

**argumentativeComponents** ^

**POST** /argumentativeComponents Returns argumentative components and annotated abstract

**relationClassification** ^

**POST** /relationClassification Returns relations between components

**picoElements** ^

**POST** /picoElements Returns PICO elements

**outcomesDetection** ^

**POST** /outcomesDetection Returns detected outcomes

**effectsPrediction** ^

**POST** /effectsPrediction Returns detected outcomes

FIGURE 5.4: Multiple screenshots to illustrate the different functionalities of ACTA and the visualization of the argument graph returned to the user.

There, the nodes are the premises and the claims automatically detected in the abstract, and the labelled edges correspond to the relations among them. In contrast to the previous version, ACTA 2.0 integrates a completely overhauled relation classification module, implementing the methods described by Mayer *et al.* [48]. Besides underlying technical changes regarding architecture, loss function and problem formulation, the most notable difference for the user is the updated linking of the arguments in the graph, which are now not only identified, but also labeled, indicating their argumentative function as either *attack* or *support*. For readability purposes, the text of the argumentative components is not shown by default in the argumentation graph. However, the user can unveil it by interacting with the graph, i.e., hovering over the respective argument component. Additionally, argument components are highlighted with different colors (evidence in blue, claims in orange) in the abstract, which is always fully shown on the right side of the window.

**PICO Element Detection.** The detected PICO elements can be visualized in a similar fashion through the *PICO Elements* button. Again, each PICO category is highlighted in a different color. For the PICO detection, we rely on the same module employed in the first version of ACTA.

**Effects on Outcomes.** As one of the major upgrades, ACTA 2.0 implements a new module to analyse the reported effects an intervention has on the outcomes (**O** of PICO) in the clinical trial abstract. Such as if an intervention increased or decreased the measured outcome, as proposed in [22]. The underlying motivation for this is twofold: first, to enrich the arguments with valuable medical information and thus increase versatility of the application; and second, to provide structured and machine-processable data, which can serve as input to a computational model of argument system [1], for instance. In the web interface of the tool, these effects can also be visualized by pressing the *Effects on Outcome* button. As a consequence, the outcomes are highlighted in the displayed abstract to the right with different colors according to their predicted effect, i.e., *Increased*, *Decreased*, *Improved*, *NoOccurrences* or *NoDifferences*.

**ACTA 2.0 Public API.** Another major upgrade is the conversion from a static monolithic pipeline to a modular and extendable system to foster versatility and re-usability. In particular, each of the processing steps, i.e., argument component detection, relation classification, PICO and effect prediction, are now independent executable units, which can be called separately via our publicly available REST API<sup>16</sup>. Researchers, developers and clinicians can now not only try them all individually, but have also the possibility to replace or add custom modules to the workflow, or build parts into their own projects. The required input and output formats for each module are defined in the documentation of the API.

<sup>16</sup><https://ns.inria.fr/acta/doc/>



**Data Format Description.** Each module takes as input a JSON file, where for the argument components, the PICO elements and Outcome Detection modules, the field “text” must be filled in with the medical text to be analyzed. For the relation classification module, the input JSON file must have the field “candidates” filled with the list of all of the argumentative components text and type (claim or premise) for which the user wants to predict the relation ( support or attack). For the effect prediction module, both the original text and the selected outcomes have to be provided in the “text” and “outcomes” fields respectively. For every module, a JSON file is produced as output with the corresponding results, either being the detected component spans or the predicted labels. All the results, including the argumentative analysis together with PICO elements and effects on outcomes, can be downloaded as a JSON file for each of the processed abstracts.

## 5.4 Open Challenges

Building usable tools that automate argument mining on real-world medical texts poses several core challenges. A primary difficulty is adapting natural language processing models to handle intricate clinical terminology and abbreviations. General pre-trained models often fail to capture the nuances of domain-specific language. Creating customized embeddings and fine-tuning in-domain corpora can improve clinical language understanding, but may require scarce expert-annotated data. Additionally, precise extraction of rhetoric structures like claims and evidence relies on accurate sequence tagging and relation classification. However, modelling the complexity of clinical reasoning patterns to robustly identify arguments is an open issue. Beyond extraction, effectively visualizing connected graphs and highlighting text semantics in an interpretable interface raises its own obstacles. Enriching arguments with contextual information like PICO elements and intervention effects provides useful analytics but complicates system design. On the implementation side, flexibility demands balancing modularity with efficient pipelines. Although systems like ACTA 2.0 make further progress, argument mining intricacies remain around domain adaptation, modelling clinical reasoning, explainable interfaces, and customizable architectures. The field would benefit from shared clinical resources and competitions to systematically address these lingering challenges. Overall, impactful medical argument mining requires overcoming issues in knowledge-aware language understanding, reasoning modelling, human-centred design, and adaptable engineering. Progress necessitates interdisciplinary collaboration across NLP, medicine, visualization, and software engineering.

## Chapter 6

# Conclusion and Future Perspectives

Assessing argument quality is critical yet challenging across domains, especially sensible fields like medicine where sound reasoning bears immense consequences. This thesis tackled key obstacles in computational argument quality assessment, both generally and for clinical applications. The core research questions aimed to advance multi-dimensional quality modelling, push boundaries in mining natural language arguments, and enable transparent assessment tailored to medical intricacies. In particular, to provide these solutions, the research questions introduced in Chapter 1 were addressed resulting in the following contributions:

**1. Modelling Argumentation Quality.** This thesis investigated the modelling of argumentation quality along dimensions like cogency, rhetoric, and reasonableness. I developed a framework to automatically assess three key quality attributes of persuasive student essays: cogency, rhetorical strategy, and reasonableness. Cogency indicates how logically sound the reasoning is, rhetorical strategy captures the writing style and use of persuasive elements, while reasonableness judges how well-balanced the essay is in considering counterarguments. To enable data-driven modelling, I built a novel corpus of 402 essays annotated by experts from the social sciences using rubrics for scoring persuasive writing.

The essays were labelled for cogency on a scale of 1 to 3 (or 0, 15, 25 in the metrics of Stapleton and Wu [19]) indicating low to high logical coherence. For rhetorical strategy, four categories were annotated reflecting different persuasive styles. Reasonableness ratings followed a similar but more fine-grained scale as cogency, with a scale of 1 to 5. I experimented with textual-only models using SVM and transformer embeddings to classify cogency and rhetorical strategy. Then I proposed a novel architecture incorporating graph-based argument structure along with text embeddings, which significantly improved predictive performance by 5-10 percentage points in macro F1 score. This demonstrated the benefits of a multi-modal approach integrating topological and linguistic information. For reasonableness, the

dataset size was insufficient for statistical modelling. Instead, I devised an algorithm leveraging cogency classifications and the argument structure graph to deterministically evaluate debate balance. Overall, this work established an encompassing framework for multi-dimensional quality assessment of argumentative writing, achieving a strong performance, with macro F1 scores of 0.77 for cogency, 0.63 for rhetorical strategy, and 0.54 for reasonableness.

**2. Tailoring argumentation quality to the medical domain.** To enable quality assessment tailored to the medical domain, this thesis introduced a specialized methodology to extract and enrich clinical argumentation. From a medical case report, I first automatically detect key entities like symptoms and test findings using natural language processing methods. These textual entities are then mapped to standardized medical ontologies like HPO to align them with validated knowledge. By linking case details to external sources, additional relevant information is obtained such as which symptoms frequently co-occur with certain diagnoses and their relative prevalence. To automatically detect the different medical entities described in the clinical cases, we experimented with different transformer-based language models such as SciBERT, BioBERT, PubMedBERT, and UmlsBERT initialized with their respective pre-trained weights specialized for the biomedical domain. We cast the symptom detection problem as a sequence tagging task, using a BIO tagging scheme. Our experiments showed that the SciBERT model achieved the best performance, with a macro F1-score of 0.86 for named entity recognition of symptoms. To accurately map the detected symptoms to the HPO medical ontology, we computed embedding vectors for each symptom and calculated the cosine distance with each HPO term to find the closest match. Our context-aware embedding approach, which summed the symptom and sentence embeddings, significantly outperformed a baseline method without context such as DASH, improving the accuracy from 0.37 to 0.53 for top-5 matches. By accounting for the contextual information, we obtained a more reliable alignment between the layperson symptom descriptions and the formal HPO terminology. Overall, the specialized neural models and context-aware embedding technique enabled effective extraction and alignment of salient clinical entities, providing a robust foundation for assessing the relevance of symptoms to potential diagnoses and generating explanatory arguments grounded in validated medical knowledge.

To further enrich this evidence, I developed a transparent prevalence function that scores the explanatory power of each symptom or finding based on medical statistics like abnormality and uniqueness [107, 111, 152–156]. Highly ranked evidence represents salient, simple reasons that should be invoked in cogent explanations. Conversely, low scores indicate extraneous details that overly complicate the reasoning if included. This evidence scoring allows for generating concise yet sound template-based explanations using only the best-ranked details. Moreover, it

enables assessing student explanations by suggesting modifications - adding high-scored evidence not utilized or removing low-scored extraneous entities. By selecting explanatory clinical details in a principled, interpretable manner, this framework provides pedagogical insights to improve clinical reasoning. Overall, it pushes quality assessment boundaries by integrating external knowledge required for validating specialized argumentation.

**3. Proof-of-Concept Applications.** Finally, this thesis made notable contributions by applying argument mining to real-world medical scenarios in Chapter 5. I collaborated on the Covid-on-the-Web project, which leveraged tools like ACTA to extract argument graphs from Coronavirus literature. These rhetorical structures augmented a knowledge graph created by disambiguating biomedical entities against resources like Wikidata. The resulting linked dataset aims to enhance evidence-based reasoning and clinician decision-making during the pandemic. Additionally, I helped develop techniques for automatically analyzing the effects of medical interventions on outcomes in clinical trials. Framing this argument mining task as a two-step pipeline of outcome extraction and then effect classification, we trained specialized transformer models to predict nuanced labels like *Increased* or *Decreased*. The pipeline achieves a macro F1-score of 0.80 for effect-on-outcome classification. For outcome extraction, a sequence tagging approach using BIO notation was employed. The extracted outcomes were then paired with their context and fed into an effect classifier implemented as a sequence classification task. By detecting outcomes in the text and categorizing their effect with specialized models, the pipeline provides structured results data to help clinicians efficiently interpret trials and synthesize findings. For example, it can be identified that "nausea decreased after taking drug A" while "liver enzyme levels increased", capturing the nuanced effects of the intervention on different outcomes. Such automated analysis facilitates evidence-based decision-making. Lastly, I upgraded the ACTA tool to version 2.0 with state-of-the-art neural models, a new module to detect intervention effects, and a modular API-based architecture. ACTA 2.0 enriches arguments with PICO elements and outcome impacts to better contextualize evidence. The public API enables customization, allowing components to be swapped within external systems. Through these applied contributions, the thesis demonstrated practical real-world benefits in adopting argument mining methods to support and enhance clinical decision making.

In brief, the research conducted in the context of this thesis showed how argument mining techniques can be specialized and applied to enhance decision making in complex real-world domains like medicine. Novel methods were introduced that model multi-faceted notions of quality, enabling computational assessment of key attributes like cogency and rhetoric. By integrating medical knowledge through contextualized concept embedding and prevalence functions, the thesis pushed the boundaries in adapting quality evaluation and argument extraction to the intricacies of clinical reasoning. These results pave the way for future work furthering this

interdisciplinary approach to augmenting clinical argumentation with validated evidence and transparent, data-driven evaluation.

## Future Perspectives

While important concepts have been carved out in my work, further research directions and future improvements are envisaged. First, larger corpora of persuasive essays are needed to explore machine learning approaches for reasonableness modelling. The current dataset lacked counterarguments, constraining available training examples for statistical models. Constructing resources with more balanced, holistic essay argumentation would enable enhanced learning of reasonableness and other qualities. Additionally, existing argument quality datasets provide only isolated argument components, not the full text with relations. Annotating arguments and relations in such data could facilitate assessment using topological structure. Finally, quality notions like cogency, though generalizable, may require adaptation or extension for certain domains. Defining new principles and corpora for specialized contexts like medicine would allow quality evaluation tailored to precise needs. Overall, the presented essay assessment framework could be expanded through richer corpora capturing variable reasonableness, annotated argument graphs, and dimensions tuned for target applications. Pursuing these directions can build on the thesis foundations to further computational quality modelling.

While the medical quality assessment techniques showed promise, enhancements could further strengthen clinical explanations and reasoning analysis. First, combining multiple knowledge bases beyond just HPO would allow a deeper validation of the claims. Advanced reasoning could also infer new symptoms from crossed ontology information. Additionally, moving beyond template-based generation to conversational explanations would enable clarifying student doubts through rule-based dialogues. Students could interactively engage the system to reinforce learning. Moreover, assessing dimension relevance and possible extensions would allow quality notions to be specialized for clinical needs. Pursuing these directions can augment the current methodology with expanded knowledge graphs, conversational pedagogical interaction, and metrics fine-tuned for healthcare. Overall, medical quality innovations could be enhanced to be even more dynamic, knowledge-rich and tailored to nuanced clinical explanation requirements.

Another promising direction is combining essay scoring techniques with argument quality assessment innovations. Existing methods automatically score student essays along dimensions like strength, coherence and organization [44, 84, 186, 187]. These could be augmented by also evaluating cogency, rhetorical strategy, and reasonableness as modelled in this thesis. Furthermore, scoring frameworks optimized for specialized domains like medicine are needed, integrating quality notions with external knowledge. Adding verified facts into the argumentation functions as

warrants in scientific argumentation, which are mostly implicitly presumed in the biomedical domain [188].

Beyond scoring, future intelligent tutoring systems could employ assessed reasons to propose argument revisions to guide students, like suggesting improvements aligned with feedback [189]. Dialogue-based interaction could also allow dynamically adapting feedback and explanations based on student needs [190]. As students revise drafts, personalized feedback conversations could address individual struggles, reinforcing successful revisions while correcting flawed reasoning.

Such personalized argumentation support would enhance learning and writing skills. By integrating automated evaluation with interactive guidance, future tools could close the assessment loop, ensuring students properly implement revision advice [186]. Alongside scoring models optimized for domains like medicine, these advances would help students construct cogent, persuasive arguments essential across education. This thesis lays the groundwork for uniting essay assessment, transparent revision analysis, and adaptive pedagogical interaction to substantially improve argumentation abilities.

Another promising direction consists of exploring different explanation strategies tailored to the student's needs and mental model. The appropriateness of arguments depends not just on logical soundness but also on aligning with the audience [3]. Strategies effective for scientific writing may differ from informal settings. Future systems could identify the student's background and adaptivity and provide personalized explanations, like balancing technical details versus intuitive analogies. Tailoring dialogue and feedback to the individual student's knowledge and abilities could enhance engagement and learning gains. Personalized learning, which includes adaptive learning, focuses on addressing the needs and goals of each student, allowing them to work at their own pace and with content tailored to their specific requirements [191]. Adaptive learning environments have also been found to have a statistically significant positive impact on student engagement [192]. This thesis assessed argument quality independently of the audience, but customizing persuasion and justification strategies based on the explainee is an important next step. Overall, the ability to discern student characteristics and adapt reasoning accordingly would allow to generate more effective explanations for each student.

Overall, this thesis addresses a few of the many facets of argumentation quality with a focus on education and the medical domain. One further future work direction is integrating the innovations from education and medical domains. Techniques like multi-modal quality classification could be combined with external knowledge alignment to enable domain-specific assessment. Expanding annotated resources with graph structure and target domains would facilitate adapting the models. Future scoring frameworks should also incorporate domain-tailored quality dimensions alongside general attributes. Additionally, personalized and conversational

systems present promising opportunities in this field. Employing revision analysis to guide students via actionable feedback would enhance writing skills. Adaptive explanations tailored to individual learners' abilities and needs could improve engagement and outcomes. With respect to medicine, this thesis provided initial steps toward evidence search, selection, appraisal and application for evidence-based practice. Significant complexity remains in the comprehensive computational modelling of clinical reasoning, necessitating substantial future work. Areas such as validating the veracity of evidence sources, weighing statistical significance, and reconciling contradictory claims require dedicated investigation. By pursuing key open challenges, future interdisciplinary research can fulfil the promise of aligning AI systems with the nuances of human argumentation.

# Bibliography

- [1] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, and S. Villata. "Towards Artificial Argumentation". In: *AI Magazine* 38.3 (2017), pp. 25–36.
- [2] J. A. Blair. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*. Vol. 21. Springer Science & Business Media, 2011.
- [3] Aristotle. *Rhetoric*. Translated by Roberts. Mineola, NY: Dover Publications, 2004.
- [4] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein. "Computational Argumentation Quality Assessment in Natural Language". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 176–187. URL: <https://www.aclweb.org/anthology/E17-1017>.
- [5] E. Saveleva, V. Petukhova, M. Mosbach, and D. Klakow. "Graph-based argument quality assessment". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 2021, pp. 1268–1280.
- [6] A. Hunter and M. Williams. "Aggregating evidence about the positive and negative effects of treatments". In: *Artificial Intelligence in Medicine* 56.3 (2012), pp. 173–190.
- [7] R. Craven, F. Toni, C. Cadar, A. Hadad, and M. Williams. "Efficient Argumentation for Medical Decision-Making". In: *Proceedings of 13th International Conference on the Principles of Knowledge Representation and Reasoning*. AAAI Press, 2012, pp. 598–602.
- [8] L. Longo and L. Hederman. "Argumentation Theory for Decision Support in Health-Care: A Comparison with Machine Learning". In: *Proceedings of the International Conference on Brain and Health Informatics 2013*. Springer. 2013, pp. 168–180.
- [9] M. Krallinger et al. "The CHEMDNER corpus of chemicals and drugs and its annotation principles". In: *Journal of cheminformatics* 7.1 (2015), pp. 1–17.
- [10] J. Li et al. "BioCreative V CDR task corpus: a resource for chemical disease relation extraction". In: *Database* 2016 (2016).



- [11] R. I. Doğan, R. Leaman, and Z. Lu. "NCBI disease corpus: a resource for disease name recognition and concept normalization". In: *Journal of biomedical informatics* 47 (2014), pp. 1–10.
- [12] L. Smith et al. "Overview of BioCreative II gene mention recognition". In: *Genome biology* 9.2 (2008), pp. 1–19.
- [13] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. "Introduction to the bio-entity recognition task at JNLPBA". In: *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Citeseer. 2004, pp. 70–75.
- [14] M. Chary, S. Parikh, A. Manini, E. Boyer, and M. Radeous. "A Review of Natural Language Processing in Medical Education". In: *Western Journal of Emergency Medicine* 20.1 (2018), pp. 78–86.
- [15] D. L. Sackett, W. M. Rosenberg, J. Gray, R. Haynes, and W. Richardson. "Evidence based medicine: What it is and what it isn't". In: *BMJ (Clinical research ed.)* 312 (1996), pp. 71–2.
- [16] G. H. Guyatt. "Evidence-based medicine". In: *ACP Journal Club* (1991), A–16.
- [17] D. L. Sackett and W. M. C. Rosenberg. "On the need for evidence-based medicine". In: *Journal of Public Health* 17.3 (1995), pp. 330–334.
- [18] L. Manchikanti. "Evidence-Based Medicine, Systematic Reviews, and Guidelines in Interventional Pain Management Part I: Introduction and General Considerations". In: *Pain physician* 11.2 (2008), pp. 161–86.
- [19] P. Stapleton and Y. ( Wu. "Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance". In: *Journal of English for Academic Purposes* 17 (2015), pp. 12–23. ISSN: 1475-1585. DOI: <https://doi.org/10.1016/j.jeap.2014.11.006>. URL: <https://www.sciencedirect.com/science/article/pii/S1475158514000824>.
- [20] S. Marro, E. Cabrio, and S. Villata. "Graph Embeddings for Argumentation Quality Assessment". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4154–4164. URL: <https://aclanthology.org/2022.findings-emnlp.306>.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Association for Computational Linguistics, 2019, pp. 4171–4186.
- [22] T. Mayer, S. Marro, E. Cabrio, and S. Villata. "Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials". In: *Artificial Intelligence in Medicine* 118 (2021), p. 102098.

- [23] P. Baroni, M. Caminada, and M. Giacomin. "An introduction to argumentation semantics". In: *The knowledge engineering review* 26.4 (2011), pp. 365–410.
- [24] T. J. Bench-Capon and P. E. Dunne. "Argumentation in artificial intelligence". In: *Artificial intelligence* 171.10-15 (2007), pp. 619–641.
- [25] M. Lippi and P. Torrioni. "Argumentation Mining: State of the Art and Emerging Trends". In: *ACM Transactions on Internet Technology* 16.2 (2016), pp. 1–25.
- [26] E. Cabrio and S. Villata. "Five Years of Argument Mining: a Data-driven Analysis". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*. International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 5427–5433.
- [27] A. Peldszus and M. Stede. "From Argument Diagrams to Argumentation Mining in Texts: A Survey". In: *International Journal of Cognitive Informatics and Natural Intelligence* 7.1 (2013), pp. 1–31.
- [28] J. Lawrence and C. Reed. "Argument Mining: A Survey". In: *Computational Linguistics* 45.4 (2020), pp. 765–818.
- [29] S. Teufel, A. Siddharthan, and C. Batchelor. "Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*. 2009, pp. 1493–1502.
- [30] J. Lawrence and C. Reed. "Argument mining: A survey". In: *Computational Linguistics* 45.4 (2020), pp. 765–818.
- [31] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed. "Automatic Detection of Arguments in Legal Texts". In: *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAAIL 2007)*. Association for Computing Machinery, 2007, pp. 225–230.
- [32] R. Mochales and M.-F. Moens. "Argumentation mining". In: *Artificial Intelligence and Law* 19.1 (2011), pp. 1–22.
- [33] E. Cabrio and S. Villata. "A Natural Language Bipolar Argumentation Approach to Support Users in Online Debate Interactions". In: *Argument & Computation* 4.3 (2013), pp. 209–230.
- [34] F. Boltužić and J. Šnajder. "Back up your Stance: Recognizing Arguments in Online Discussions". In: *Proceedings of the 1st Workshop on Argumentation Mining (ArgMining 2014)*. Association for Computational Linguistics, 2014, pp. 49–58.
- [35] J. Daxenberger, S. Eger, I. Habernal, C. Stab, and I. Gurevych. "What is the Essence of a Claim? Cross-Domain Claim Identification". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by M. Palmer, R. Hwa, and S. Riedel. Copenhagen, Denmark: Association for

- Computational Linguistics, Sept. 2017, pp. 2055–2066. DOI: 10 . 18653 / v1 / D17-1218. URL: <https://aclanthology.org/D17-1218>.
- [36] P. Baroni, M. Caminada, and M. Giacomin. “An introduction to argumentation semantics”. In: *Knowl. Eng. Rev.* 26.4 (2011), pp. 365–410. DOI: 10 . 1017 / S0269888911000166. URL: <https://doi.org/10.1017/S0269888911000166>.
- [37] R. H. Johnson and J. A. Blair. *Logical self-defense*. Idea, 2006.
- [38] F. H. Van Eemeren and R. Grootendorst. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press, 2004.
- [39] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein. “Computational argumentation quality assessment in natural language”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 2017, pp. 176–187.
- [40] H. Wachsmuth, B. Stein, and Y. Ajjour. ““PageRank” for Argument Relevance”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1117–1127. URL: <https://aclanthology.org/E17-1105>.
- [41] I. Habernal and I. Gurevych. “Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1589–1599.
- [42] A. Toledo, S. Gretz, E. Cohen-Karlik, R. Friedman, E. Venezian, D. Lahav, M. Jacovi, R. Aharonov, and N. Slonim. “Automatic Argument Quality Assessment - New Datasets and Methods”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5625–5635. DOI: 10 . 18653 / v1 / D19 - 1564. URL: <https://aclanthology.org/D19-1564>.
- [43] Z. Rahimi, D. J. Litman, R. Correnti, L. C. Matsumura, E. Wang, and Z. Kisa. “Automatic scoring of an analytical response-to-text assessment”. In: *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12*. Springer. 2014, pp. 601–610.
- [44] I. Persing, A. Davis, and V. Ng. “Modeling organization in student essays”. In: *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010, pp. 229–239.

- [45] L. Ng, A. Lauscher, J. Tetreault, and C. Napoles. "Creating a Domain-diverse Corpus for Theory-based Argument Quality Assessment". In: *Proceedings of the 7th Workshop on Argument Mining*. Ed. by E. Cabrio and S. Villata. Online: Association for Computational Linguistics, Dec. 2020, pp. 117–126. URL: <https://aclanthology.org/2020.argmining-1.13>.
- [46] I. Persing and V. Ng. "Lightly-supervised modeling of argument persuasiveness". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2017, pp. 594–604.
- [47] S. Gretz, R. Friedman, E. Cohen-Karlik, A. Toledo, D. Lahav, R. Aharonov, and N. Slonim. "A large-scale dataset for argument quality ranking: Construction and analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7805–7813.
- [48] T. Mayer, S. Marro, E. Cabrio, and S. Villata. "Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials". In: *Artificial Intelligence in Medicine* (2021), p. 102098.
- [49] O. Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology". In: *Nucleic acids research* 32.suppl\_1 (2004), pp. D267–D270.
- [50] S. Köhler et al. "The human phenotype ontology in 2017". In: *Nucleic acids research* 45.D1 (2017), pp. D865–D876.
- [51] K. S. Jones. "Natural language processing: a historical review". In: *University of Cambridge* (2001), pp. 2–10.
- [52] N. Indurkha and F. J. Damerau. *Handbook of natural language processing*. Vol. 2. CRC Press, 2010.
- [53] W. J. Hutchins. "Machine translation: A brief history". In: *Concise history of the language sciences*. Elsevier, 1995, pp. 431–445.
- [54] N. Chomsky. *On nature and language*. Cambridge University Press, 2002.
- [55] S. T. Piantadosi, H. Tily, and E. Gibson. "The communicative function of ambiguity in language". In: *Cognition* 122.3 (2012), pp. 280–291.
- [56] A. Neelakantan, J. Shankar, A. Passos, and A. McCallum. "Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by A. Moschitti, B. Pang, and W. Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1059–1069. DOI: 10.3115/v1/D14-1113. URL: <https://aclanthology.org/D14-1113>.
- [57] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". In: *Workshop Track Proceedings of the 1st International Conference on Learning Representations (ICLR 2013)*. 2013.

- [58] J. Pennington, R. Socher, and C. D. Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Association for Computational Linguistics, 2014, pp. 1532–1543.
- [59] V. Shwartz and I. Dagan. "Still a pain in the neck: Evaluating text representations on lexical composition". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 403–419.
- [60] S. Nair, M. Srinivasan, and S. Meylan. "Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge". In: *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*. Online: Association for Computational Linguistics, Dec. 2020, pp. 129–141. URL: <https://aclanthology.org/2020.cogalex-1.16>.
- [61] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*. Association for Computational Linguistics, 2018, pp. 2227–2237.
- [62] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.
- [63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*. Curran Associates, Inc., 2017, pp. 6000–6010.
- [64] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. "Improving language understanding by generative pre-training". In: *OpenAI report* (2018).
- [65] W. L. Taylor. "'Cloze Procedure': A New Tool for Measuring Readability". In: *Journalism Quarterly* 30.4 (1953), pp. 415–433.
- [66] I. Beltagy, K. Lo, and A. Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 2019, pp. 3615–3620.
- [67] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.
- [68] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Trans. Comput. Healthcare* 3.1 (Oct. 2021). DOI: 10.1145/3458754. URL: <https://doi.org/10.1145/3458754>.

- [69] Y. Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019).
- [70] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: <https://openreview.net/forum?id=H1eA7AEtvS>.
- [71] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: <https://aclanthology.org/D19-1410>.
- [72] S. Marro, E. Cabrio, and S. Villata. "Argumentation Quality Assessment: an Argument Mining Approach". In: *ECA 2022 - European conference on argumentation*. Rome, Italy, Oct. 2022. URL: <https://hal.science/hal-03934466>.
- [73] G. R. Simari and I. Rahwan, eds. *Argumentation in Artificial Intelligence*. Springer, 2009. ISBN: 978-0-387-98196-3. DOI: 10.1007/978-0-387-98197-0. URL: <https://doi.org/10.1007/978-0-387-98197-0>.
- [74] A. Hunter. "Argument Strength in Probabilistic Argumentation Using Confirmation Theory". In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21-24, 2021, Proceedings*. Ed. by J. Vejnárová and N. Wilson. Vol. 12897. Lecture Notes in Computer Science. Springer, 2021, pp. 74–88. DOI: 10.1007/978-3-030-86772-0\_6. URL: [https://doi.org/10.1007/978-3-030-86772-0\\_6](https://doi.org/10.1007/978-3-030-86772-0_6).
- [75] L. Amgoud, D. Doder, and S. Vesic. "Evaluation of argument strength in attack graphs: Foundations and semantics". In: *Artif. Intell.* 302 (2022), p. 103607. DOI: 10.1016/j.artint.2021.103607. URL: <https://doi.org/10.1016/j.artint.2021.103607>.
- [76] C. da Costa Pereira, A. Tettamanzi, and S. Villata. "Changing One's Mind: Erase or Rewind?" In: *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*. Ed. by T. Walsh. IJCAI/AAAI, 2011, pp. 164–171. DOI: 10.5591/978-1-57735-516-8/IJCAI11-039. URL: <https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-039>.
- [77] T. J. M. Bench-Capon. "Persuasion in Practical Argument Using Value-based Argumentation Frameworks". In: *J. Log. Comput.* 13.3 (2003), pp. 429–448.

- DOI: 10.1093/logcom/13.3.429. URL: <https://doi.org/10.1093/logcom/13.3.429>.
- [78] H. Wachsmuth and T. Werner. “Intrinsic Quality Assessment of Arguments”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 6739–6745.
- [79] J. Lawrence and C. Reed. “Argument Mining: A Survey”. In: *Comput. Linguistics* 45.4 (2019), pp. 765–818. DOI: 10.1162/coli\_a\_00364. URL: [https://doi.org/10.1162/coli%5C\\_a%5C\\_00364](https://doi.org/10.1162/coli%5C_a%5C_00364).
- [80] A. Lauscher, H. Wachsmuth, I. Gurevych, and G. Glavaš. “Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation”. In: *Transactions of the Association for Computational Linguistics* 10 (2022). Ed. by B. Roark and A. Nenkova, pp. 1392–1422. DOI: 10.1162/tacl\_a\_00525. URL: <https://aclanthology.org/2022.tacl-1.80>.
- [81] C. Stab and I. Gurevych. “Parsing Argumentation Structures in Persuasive Essays”. In: *Computational Linguistics* 43.3 (Sept. 2017), pp. 619–659. DOI: 10.1162/coli\_a\_00295. URL: <https://www.aclweb.org/anthology/J17-3005>.
- [82] H. Wachsmuth, B. Stein, and Y. Ajour. “PageRank for argument relevance”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics, 2017, pp. 1117–1127.
- [83] I. Persing and V. Ng. “Modeling thesis clarity in student essays”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 260–269.
- [84] I. Persing and V. Ng. “Modeling argument strength in student essays”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 543–552.
- [85] W. Luo and D. Litman. “Determining the quality of a student reflective response”. In: *The twenty-ninth international FLAIRS Conference*. 2016.
- [86] A. Lauscher, L. Ng, C. Napoles, and J. Tetreault. “Rhetoric, Logic, and Dialectic: Advancing Theory-based Argument Quality Assessment in Natural Language Processing”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4563–4574. URL: <https://www.aclweb.org/anthology/2020.coling-main.402>.
- [87] R. Duthie, K. Budzynska, and C. Reed. “Mining ethos in political debate”. In: *Computational Models of Argument: Proceedings from the Sixth International Conference on Computational Models of Argument (COMMA)*. IOS Press. 2016, pp. 299–310.

- [88] H. Zhang and D. Litman. "Essay Quality Signals as Weak Supervision for Source-based Essay Scoring". In: *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*. Online: Association for Computational Linguistics, Apr. 2021, pp. 85–96. URL: <https://aclanthology.org/2021.bea-1.9>.
- [89] L. Coertjens, M. Lesterhuis, S. Verhavert, R. Van Gasse, and S. De Maeyer. "Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering= Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment". In: *Pedagogische studiën* 94.4 (2017), pp. 283–303.
- [90] M. Sasaki and K. Hirose. "Development of an analytic rating scale for Japanese L1 writing". In: *Language Testing* 16.4 (1999), pp. 457–478.
- [91] B. Rozemberczki and R. Sarkar. "Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 1325–1334. ISBN: 9781450368599. DOI: 10.1145/3340531.3411866. URL: <https://doi.org/10.1145/3340531.3411866>.
- [92] I. Beltagy, M. E. Peters, and A. Cohan. "Longformer: The Long-Document Transformer". In: *CoRR abs/2004.05150* (2020). arXiv: 2004.05150. URL: <https://arxiv.org/abs/2004.05150>.
- [93] W. Medhat, A. Hassan, and H. Korashy. "Sentiment analysis algorithms and applications: A survey". In: *Ain Shams engineering journal* 5.4 (2014), pp. 1093–1113.
- [94] S. Zad, M. Heidari, H. James Jr, and O. Uzuner. "Emotion detection of textual data: An interdisciplinary survey". In: *2021 IEEE World AI IoT Congress (AllIoT)*. IEEE, 2021, pp. 0255–0261.
- [95] N. Keivandarian and M. Carvalho. "A Survey on Sentiment Classification Methods and Challenges". In: *The International FLAIRS Conference Proceedings*. Vol. 36. 2023.
- [96] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.
- [97] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen. "CARER: Contextualized Affect Representations for Emotion Recognition". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3687–3697. DOI: 10.18653/v1/D18-1404. URL: <https://www.aclweb.org/anthology/D18-1404>.



- [98] C.-C. Chang and C.-J. Lin. "LIBSVM: a library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27.
- [99] A. Cutler, D. R. Cutler, and J. R. Stevens. "Random forests". In: *Ensemble machine learning*. Springer, 2012, pp. 157–175.
- [100] R. Panchendrarajan and A. Amaresan. "Bidirectional LSTM-CRF for Named Entity Recognition". In: *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Ed. by S. Politzer-Ahles, Y.-Y. Hsu, C.-R. Huang, and Y. Yao. Hong Kong: Association for Computational Linguistics, Jan. 2018. URL: <https://aclanthology.org/Y18-1061>.
- [101] A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [102] T. Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Q. Liu and D. Schlangen. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: <https://aclanthology.org/2020.emnlp-demos.6>.
- [103] B. Rozemberczki, O. Kiss, and R. Sarkar. "Karate Club: An API Oriented Open-source Python Framework for Unsupervised Learning on Graphs". In: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*. ACM, 2020, pp. 3125–3132.
- [104] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [105] P.-L. Chuang and X. Yan. "An investigation of the relationship between argument structure and essay quality in assessed writing". In: *Journal of Second Language Writing* 56 (2022), p. 100892.
- [106] S. Marro, B. Molinet, E. Cabrio, and S. Villata. "Natural Language Explanatory Arguments for Correct and Incorrect Diagnoses of Clinical Cases". In: *ICAART 2023-15th International Conference on Agents and Artificial Intelligence*. Vol. 1. 2023, pp. 438–449.
- [107] T. Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.
- [108] T. Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artif. Intell.* 267 (2019), pp. 1–38. DOI: 10.1016/j.artint.2018.07.007. URL: <https://doi.org/10.1016/j.artint.2018.07.007>.

- [109] G. H. Harman. "The inference to the best explanation". In: *The philosophical review* 74.1 (1965), pp. 88–95.
- [110] T. Trabasso and J. Bartolone. "Story understanding and counterfactual reasoning." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.5 (2003), p. 904.
- [111] D. J. Hilton and B. R. Slugoski. "Knowledge-based causal attribution: The abnormal conditions focus model." In: *Psychological review* 93.1 (1986), p. 75.
- [112] T. Mayer, S. Marro, E. Cabrio, and S. Villata. "Generating Adversarial Examples for Topic-Dependent Argument Classification". In: *Proceedings of 8th International Conference on Computational Models of Argument (COMMA 2020)*. IOS Press, 2020, pp. 33–44.
- [113] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. "ERNIE: Enhanced Language Representation with Informative Entities". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, 2019, pp. 1441–1451.
- [114] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang. "ERNIE 2.0: A Continual Pre-training Framework for Language Understanding". In: *CoRR abs/1907.12412* (2019).
- [115] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou. "Semantics-aware BERT for language understanding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 9628–9635.
- [116] R. Speer, J. Chin, and C. Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.
- [117] K. Faldu, A. Sheth, P. Kikani, and H. Akbari. "Ki-bert: Infusing knowledge context for better language and domain understanding". In: *arXiv preprint arXiv:2104.08145* (2021).
- [118] W. Zhou, D.-H. Lee, R. K. Selvam, S. Lee, B. Y. Lin, and X. Ren. "Pre-training text-to-text transformers for concept-centric common sense". In: *arXiv preprint arXiv:2011.07956* (2020).
- [119] Y.-L. Tuan, S. Beygi, M. Fazel-Zarandi, Q. Gao, A. Cervone, and W. Y. Wang. "Towards Large-Scale Interpretable Knowledge Graph Reasoning for Dialogue Systems". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 383–395. DOI: 10.18653/v1/2022.findings-acl.33. URL: <https://aclanthology.org/2022.findings-acl.33>.
- [120] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

- [121] F. K. Došilović, M. Brčić, and N. Hlupić. “Explainable artificial intelligence: A survey”. In: *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. 2018, pp. 0210–0215.
- [122] E. Reiter. “Natural Language Generation Challenges for Explainable AI”. In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. Ed. by J. M. Alonso and A. Catala. Association for Computational Linguistics, 2019, pp. 3–7. DOI: 10.18653/v1/W19-8402. URL: <https://aclanthology.org/W19-8402>.
- [123] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, and V. Patkar. “Argumentation-based inference and decision making—A medical perspective”. In: *IEEE intelligent systems* 22.6 (2007), pp. 34–41.
- [124] J. Marques-Silva and A. Ignatiev. “Delivering Trustworthy AI through formal XAI”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2022, pp. 12342–12350.
- [125] S. Köhler et al. “The human phenotype ontology in 2021”. In: *Nucleic acids research* 49.D1 (2021), pp. D1207–D1217.
- [126] Y. Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).
- [127] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”. In: *ICLR*. 2020. URL: <https://openreview.net/pdf?id=r1xMH1BtvB>.
- [128] M. Honnibal and I. Montani. “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. To appear. 2017.
- [129] H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson. “Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python”. In: *AMIA Annual Symposium Proceedings*. Vol. 2021. American Medical Informatics Association. 2021, p. 438.
- [130] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu. “CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines”. In: *Journal of the American Medical Informatics Association* 25.3 (2018), pp. 331–336.
- [131] U. Naseem, M. Khushi, V. B. Reddy, S. Rajendran, I. Razzak, and J. Kim. “BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition”. In: *2021 International Joint Conference on Neural Networks (IJCNN)* (2021), pp. 1–7.
- [132] K. raj Kanakarajan, B. Kundumani, and M. Sankarasubbu. “BioELECTRA: pretrained biomedical text encoder using discriminators”. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. 2021, pp. 143–154.

- [133] I. Beltagy, K. Lo, and A. Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. DOI: 10.18653/v1/D19-1371. URL: <https://aclanthology.org/D19-1371>.
- [134] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.
- [135] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *Proceedings of ACL*. 2020.
- [136] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong. “UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Online: Association for Computational Linguistics, June 2021, pp. 1744–1753. DOI: 10.18653/v1/2021.naacl-main.139. URL: <https://aclanthology.org/2021.naacl-main.139>.
- [137] H. Ngai and F. Rudzicz. “Doctor XAvIer: Explainable Diagnosis on Physician-Patient Dialogues and XAI Evaluation”. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 337–344. DOI: 10.18653/v1/2022.bionlp-1.33. URL: <https://aclanthology.org/2022.bionlp-1.33>.
- [138] E. Manzini, J. Garrido-Aguirre, J. Fonollosa, and A. Perera-Lluna. “Mapping layperson medical terminology into the Human Phenotype Ontology using neural machine translation models”. In: *Expert Systems with Applications* 204 (2022), p. 117446.
- [139] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. “e-SNLI: Natural Language Inference with Natural Language Explanations”. In: *NeurIPS*. 2018.
- [140] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. “A large annotated corpus for learning natural language inference”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by L. Màrquez, C. Callison-Burch, and J. Su. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: <https://aclanthology.org/D15-1075>.

- [141] S. Kumar and P. Talukdar. "NILE : Natural Language Inference with Faithful Natural Language Explanations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 8730–8742. DOI: 10 . 18653 / v1 / 2020 . acl - main . 771. URL: <https://aclanthology.org/2020.acl-main.771>.
- [142] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.
- [143] S. Narang, C. Raffel, K. Lee, A. Roberts, N. Fiedel, and K. Malkan. "Wt5?! training text-to-text models to explain their predictions". In: *arXiv preprint arXiv:2004.14546* (2020).
- [144] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *J. Mach. Learn. Res.* 21.1 (Jan. 2020). ISSN: 1532-4435.
- [145] J. R. Josephson and S. G. Josephson. *Abductive inference: Computation, Philosophy, Technology*. Cambridge University Press, 1994.
- [146] D. G. Campos. "On the distinction between Peirce's abduction and Lipton's inference to the best explanation". In: *Synthese* 180.3 (2011), pp. 419–442.
- [147] S. Dragulinescu. "Inference to the best explanation and mechanisms in medicine". In: *Theoretical medicine and bioethics* 37 (3 2016), pp. 211–232.
- [148] E. Reiter and R. Dale. "Building applied natural language generation systems". In: *Natural Language Engineering* 3.1 (1997), pp. 57–87.
- [149] A. Abujabal, R. S. Roy, M. Yahya, and G. Weikum. "Quint: Interpretable question answering over knowledge bases". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2017, pp. 61–66.
- [150] D. Hilton. "Social attribution and explanation". In: Jan. 2016.
- [151] G. Hesslow. "The problem of causal selection". In: *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988), pp. 11–32.
- [152] B. Rehder. "A causal-model theory of conceptual representation and categorization." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.6 (2003), p. 1141.
- [153] B. Rehder. "When similarity and causality compete in category-based property generalization". In: *Memory & Cognition* 34 (2006), pp. 3–16.
- [154] J. L. McClure, R. M. Sutton, and D. J. Hilton. "Implicit and explicit processes in social judgments: The role of goal-based explanations." In: *Social judgments: Implicit and explicit processes* 5 (2003), p. 306.

- [155] J. Samland and M. R. Waldmann. “Do social norms influence causal inferences?” In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36. 2014.
- [156] D. J. Hilton and L. M. John. “The course of events: counterfactuals, causal sequences, and explanation”. In: *The psychology of counterfactual thinking*. Routledge, 2007, pp. 56–72.
- [157] A. Holzinger, A. Carrington, and H. Müller. “Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations”. In: *KI-Künstliche Intelligenz* 34.2 (2020), pp. 193–198.
- [158] C. Panigutti, A. Perotti, and D. Pedreschi. “Doctor XAI: an ontology-based approach to black-box sequential data classification explanations”. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 629–639.
- [159] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger. “Explainable AI: the new 42?” In: *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*. Springer. 2018, pp. 295–303.
- [160] T. Gall et al. “Defining disease, diagnosis, and translational medicine within a homeostatic perturbation paradigm: The national institutes of health undiagnosed diseases program experience”. In: *Frontiers in medicine* 4 (2017), p. 62.
- [161] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. “What disease does this patient have? a large-scale open domain question answering dataset from medical exams”. In: *Applied Sciences* 11.14 (2021), p. 6421.
- [162] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, and A. Moreno-Sandoval. “A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine”. In: *BMC medical informatics and decision making* 21.1 (2021), pp. 1–19.
- [163] D. Albright et al. “Towards comprehensive syntactic and semantic annotations of the clinical narrative”. In: *Journal of the American Medical Informatics Association* 20.5 (2013), pp. 922–930.
- [164] S. Mohan and D. Li. “Medmentions: A large biomedical corpus annotated with umls concepts”. In: *arXiv preprint arXiv:1902.09476* (2019).
- [165] J. L. Fleiss. “Measuring nominal scale agreement among many raters”. In: *Psychological bulletin* 76.5 (1971), pp. 378–382.
- [166] P. Deka, A. Jurek-Loughrey, and P. Deepak. “Improved Methods To Aid Unsupervised Evidence-based Fact Checking For Online Health News”. In: *Journal of Data Intelligence* 3.4 (2022), pp. 474–504.

- [167] R. Aggeri et al. *HiTZ@Antidote: Argumentation-driven Explainable Artificial Intelligence for Digital Medicine*. 2023.
- [168] J. Lu, J. Li, B. Wallace, Y. He, and G. Pergola. “NapSS: Paragraph-level Medical Text Simplification via Narrative Prompting and Sentence-matching Summarization”. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1079–1091. URL: <https://aclanthology.org/2023.findings-eacl.80>.
- [169] P. Deka, A. Jurek-Loughrey, et al. “Evidence Extraction to Validate Medical Claims in Fake News Detection”. In: *International Conference on Health Information Science*. Springer. 2022, pp. 3–15.
- [170] F. Michel et al. “Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research”. In: *Proceedings of the 19th International Semantic Web Conference (ISWC 2020)*. in-press, 2020.
- [171] B. Molinet, S. Marro, E. Cabrio, S. Villata, and T. Mayer. “ACTA 2.0: A Modular Architecture for Multi-Layer Argumentative Analysis of Clinical Trials”. In: *IJCAI 2022-Thirty-First International Joint Conference on Artificial Intelligence*. 2022.
- [172] T. Mayer, E. Cabrio, and S. Villata. “ACTA A Tool for Argumentative Clinical Trial Analysis”. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*. International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6551–6553.
- [173] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. “Querying the Semantic Web with Corese Search Engine”. In: *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*. IOS Press, 2004, pp. 705–709.
- [174] R. A. Cava, C. M. D. S. Freitas, and M. Winckler. “ClusterVis: visualizing nodes attributes in multivariate graphs”. In: *Proceedings of the 32nd Symposium on Applied Computing (SAC 2017)*. ACM, 2017, pp. 174–179.
- [175] L. L. Wang et al. “CORD-19: The COVID-19 Open Research Dataset”. In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Ed. by K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, and B. Wallace. Online: Association for Computational Linguistics, July 2020. URL: <https://aclanthology.org/2020.nlpCOVID19-acl.1>.
- [176] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. “Improving efficiency and accuracy in multilingual entity extraction”. In: *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTiCS 2013)*. Association for Computing Machinery, 2013, pp. 121–124.
- [177] C. Jonquet, N. H. Shah, and M. A. Musen. “The open biomedical annotator”. In: *Summit on Translational Bioinformatics 2009 (2009)*, p. 56.

- [178] X. Wang et al. "EVIDENCEMINER: Textual Evidence Discovery for Life Sciences". In: *Proceedings of ACL 2022: System Demonstrations*. 2020, pp. 56–62.
- [179] I. J. Marshall, J. Kuiper, and B. C. Wallace. "RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials". In: *Journal of the American Medical Informatics Association* 23.1 (2016), pp. 193–201.
- [180] I. Marshall, J. Kuiper, E. Banner, and B. C. Wallace. "Automating Biomedical Evidence Synthesis: RobotReviewer". In: *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, 2017, pp. 7–12.
- [181] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, and I. Sim. "ExaCT: Automatic extraction of clinical trial characteristics from journal publications". In: *BMC medical informatics and decision making* 10 (2010), p. 56.
- [182] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace. "Inferring Which Medical Treatments Work from Reports of Clinical Trials". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Association for Computational Linguistics, 2019, pp. 3705–3717.
- [183] A. Dhrangadhariya, G. Aguilar, T. Solorio, R. Hilfiker, and H. Müller. "End-to-End Fine-Grained Neural Entity Recognition of Patients, Interventions, Outcomes". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. 2021, pp. 65–77.
- [184] D. Jin and P. Szolovits. "PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks". In: *Proceedings of the 17th Workshop on Biomedical Natural Language Processing (BioNLP 2018)*. 2018, pp. 67–75.
- [185] A. Trenta, A. Hunter, and S. Riedel. "Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints". In: *CoRR* abs/1509.05209 (2015).
- [186] M. D. Shermis and J. Burstein. "Handbook of automated essay evaluation: Current applications and new directions". In: (2013).
- [187] Y. Attali and J. Burstein. "Automated Essay Scoring With e-rater® V.2". In: *The Journal of Technology, Learning and Assessment* 4.3 (Feb. 2006). URL: <https://ejournals.bc.edu/index.php/jtla/article/view/1650>.
- [188] N. L. Green. "Towards mining scientific discourse using argumentation schemes". In: *Argument & Computation* 9.2 (2018), pp. 121–135.
- [189] T. Afrin, E. L. Wang, D. Litman, L. C. Matsumura, and R. Correnti. "Annotation and Classification of Evidence and Reasoning Revisions in Argumentative Writing". In: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA → Online: Association for Computational Linguistics, July 2020, pp. 75–84. DOI: 10.18653/v1/2020.bea-1.7. URL: <https://aclanthology.org/2020.bea-1.7>.



- 
- [190] V. Rus, M. Lintean, and R. Azevedo. "Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor." In: *International working group on educational data mining* (2009).
- [191] J. F. Pane, E. D. Steiner, M. D. Baird, L. S. Hamilton, and J. D. Pane. *How Does Personalized Learning Affect Student Achievement?* Santa Monica, CA: RAND Corporation, 2017. DOI: 10.7249/RB9994.
- [192] H. A. El-Sabagh. "Adaptive e-learning environment based on learning styles and its impact on development students' engagement". In: *International Journal of Educational Technology in Higher Education* 18.1 (2021), pp. 1–24.