Analysing and Unveiling Argumentative Fallacies in Political Debates. A use case on the U.S. Presidential Debates from 1960 to 2020

Pierpaolo Goffredo

Analyser et dévoiler l'argumentation fallacieuse dans les débats politiques. Le cas d'usage des débats présidentiels américains de 1960 à 2020

Pierpaolo GOFFREDO

Université Côte d'Azur, Inria, CNRS, I3S

Présentée en vue de l'obtention du grade de docteur en Informatique **d'**Université Côte d'Azur

Dirigée par : Serena VILLATA, Directrice de Recherche, Université Côte d'Azur

Co-dirigée par : Elena CABRIO, Professeur des Universités, Université Côte d'Azur

Soutenue le : 16 Décembre 2024

Devant le jury, composé de :

Président du Jury : Fabien GANDON, Directeur de Recherche, INRIA

Rapporteurs :

Smaranda MURESAN, Professeur des Universités, Columbia University

Henning WACHSMUTH, Professeur des Universités, Leibniz University Hannover

Examinateur :

Anthony HUNTER, Professeur des Universités, University College London

UNIVERSITÉ CÔTE D'AZUR Inria cnrs i3S sophia antipolis

UNIVERSITÉ CÔTE D'AZUR

Analyser et dévoiler l'argumentation fallacieuse dans les débats politiques. Le cas d'usage des débats présidentiels américains de 1960 à 2020

Analysing and Unveiling Argumentative Fallacies in Political Debates. A use case on the U.S. Presidential Debates from 1960 to 2020

COMPOSITION DU JURY

President of the Jury:
Fabien GANDON Directeur de Recherche, INRIA

Rapporteurs:
Smaranda MURESAN Professeur des Universités, Columbia University
Henning WACHSMUTH Professeur des Universités, Leibniz University Hannover

Examinateur:
Anthony HUNTER Professeur des Universités, University College London

Directrice de thèse:
Serena VILLATA Directrice de Recherche, Université Côte d'Azur, Inria, CNRS, I3S

Co-Directrice de thèse:
Elena CABRIO Professeur des Universités, Université Côte d'Azur, Inria, CNRS, I3S

December 16, 2024

Abstract

This thesis presents a comprehensive study on the automated analysis of argumentative structures and fallacious arguments in political debates, with a particular focus on the U.S. presidential campaigns from 1960 to 2020. This research addresses the critical challenge of combating disinformation and propagandist content in political discourse, highlighting the nefarious effects of misleading arguments on citizens and policymakers. A first contribution of this thesis is the curation of the ElecDeb60to20 dataset, an extensive resource comprising 44 U.S. presidential debates. This dataset is annotated with 55,679 argument components (29,624 claims and 26,055 premises), 25,524 argumentative relations (3,835 attacks and 21,689 supports), and 1,640 fallacious arguments across six categories, i.e., Ad Hominem, Appeal To Authority, Appeal To Emotion, False Cause, Slogan, and Slippery Slope. Building upon this extensive resource with different annotation layers (i.e., argument components, relations, fallacy classes), the second contribution of this thesis consists on the development of a full argument mining pipeline for political debates. This pipeline achieves state-of-the-art performance, with an average F1 Score of 47% for component detection and 69% for relation prediction. The approaches developed are then employed to deploy DISPUTool 2.0, a tool for extracting argument components and relations from political debates provided by the user. The research focus of the thesis then pivots on the critical issue of fallacies in political argumentation. The first major contribution in this area is the development and evaluation of different neural architectures based on transformers for the automatic classification of fallacious arguments among the six identified categories. This approach demonstrated significant improvements over baseline methods, achieving an F1 Score of 84% for fallacy classification. Further refinements led to a model capable of simultaneously identifying and classifying fallacies, which yielded an F1 Score of 74% for detection and for classification, enhancing both the efficiency and accuracy of the process. Finally, the last contribution of the thesis tackles the challenge of "unveiling" fallacies — not just detecting fallacious arguments, but also reformulating them into non-fallacious arguments. This experimental setting involved leveraging Large Language Models (LLMs) through carefully designed prompting strategies. A novel evaluation methodology has been developed to assess the quality of these generated non-fallacious arguments, including user studies to gauge their Relevance, Suitability, and Cogency.

In conclusion, this research focused on developing models and tools to promote a more informed and healthier political discourse. By automating the identification, classification, and reformulation of fallacious arguments, this work contributes to the broader goal of mitigating the spread of disinformation and propaganda in political debates. The findings have significant implications for critical thinking education, the

development of contemporary argumentation technologies, and the ongoing effort to foster a more robust and truthful political dialogue in democratic societies.

Keywords: Argument Mining, Political Debates, Fallacy Detection

Résumé

Cette thèse présente une étude complète sur l'analyse automatisée des structures argumentatives et des arguments fallacieux dans les débats politiques, avec un accent particulier sur les campagnes présidentielles américaines de 1960 à 2020. Cette recherche aborde le défi critique de la lutte contre la désinformation et le contenu propagandiste dans le discours politique, en soulignant les effets néfastes des arguments trompeurs sur les citoyens et les décideurs. La première contribution de cette thèse est la curation de l'ensemble de données ElecDeb60to20, une ressource étendue comprenant 44 débats présidentiels américains. Cet jeu de données est annoté avec 55 679 composants d'arguments (29 624 conclusions et 26 055 prémisses), 25 524 relations argumentatives (3 835 attaques et 21 689 supports), et 1 640 arguments fallacieux dans six catégories, à savoir Ad Hominem, Appeal To Authority, Appeal To Emotion, False Cause, Slogan et Slippery Slope. En s'appuyant sur cette ressource étendue avec différentes couches d'annotation (c'est-à-dire les composants des arguments, les relations, les classes d'arguments fallacieux), la deuxième contribution de cette thèse consiste à développer un pipeline complet d'extraction d'arguments pour les débats politiques. Ce pipeline atteint des performances satisfaisantes, avec un score F1 moyen de 47% pour la détection des composants et de 69% pour la prédiction des relations. Les approches développées sont ensuite utilisées pour déployer DISPUTool 2.0, un outil d'extraction de composants et de relations d'arguments à partir de débats politiques fournis par l'utilisateur. L'axe de recherche de la thèse s'articule ensuite autour de la question cruciale des arguments fallacieux dans l'argumentation politique. La première contribution majeure dans ce domaine est le développement et l'évaluation de différentes architectures neuronales basées sur des transformateurs pour la classification automatique des arguments fallacieux parmi les six catégories identifiées. Cette approche a démontré des améliorations significatives par rapport aux méthodes de base, atteignant un score F1 de 84% pour la classification des arguments fallacieux. D'autres améliorations ont conduit à un modèle capable d'identifier et de classer simultanément les arguments fallacieux, ce qui a permis d'obtenir un score F1 de 74% pour la détection et la classification, améliorant ainsi à la fois l'efficacité et la précision du processus. Enfin, la dernière contribution de la thèse s'attaque au défi de "dévoiler" les fallacies — non seulement en détectant les arguments fallacieux, mais aussi en les reformulant en arguments non fallacieux. Ce cadre expérimental implique l'utilisation de grands modèles de langage (LLM) par le biais de stratégies d'incitation. Une nouvelle méthodologie d'évaluation a été mise au point pour évaluer la qualité des arguments non fallacieux générés, y compris des études avec les utilisateurs, pour mesurer leur pertinence, leur adéquation et leur cohérence.

En conclusion, cette recherche s'est concentrée sur le développement de modèles

et d'outils visant à promouvoir un discours politique plus informé et sain. En automatisant l'identification, la classification et la reformulation des arguments fallacieux, ce travail contribue à l'objectif plus large d'atténuer la propagation de la désinformation et de la propagande dans les débats politiques. Les résultats ont des implications significatives pour le développement de la pensée critique à travers la proposition de technologies d'argumentation conçues pour favoriser un dialogue politique plus robuste et transparent dans la société démocratique.

Mots clés: Extraction d'arguments, Débats politiques, Identification d'Arguments Fallacieux

Acknowledgements

As I conclude this significant chapter of my academic journey, I would like to express my deepest gratitude to all those who have played a crucial role in the realization of this doctoral thesis. First and foremost, I extend my thanks to my supervisors, Professor Serena and Professor Elena. Your support, expert guidance, and profound insights have been instrumental in shaping not only this research but also my growth as a scholar. Your patience, encouragement, and constructive feedback throughout this journey have been invaluable.

I am grateful to the reviewers and members of the jury who have dedicated their time and expertise to evaluate this work. Your thorough examination and constructive criticisms have significantly enhanced the quality of this thesis.

My sincere appreciation goes to the WIMMICS team for providing a nurturing and stimulating environment for my research over these years. The collaborative atmosphere, state-of-the-art facilities, and the opportunity to engage with brilliant minds have been crucial to the success of this project. The seminars, workshops, and informal discussions with fellow researchers have broadened my horizons and contributed immensely to my academic development.

I would like to express my gratitude to 3IA for their financial support through my doctoral fellowship. Moreover, I am particularly thankful for the opportunity 3IA provided to engage in teaching activities. These experiences have not only been extremely formative and enriching but have also allowed me to develop valuable skills in communicating complex ideas and mentoring students.

A special mention goes to my colleagues and fellow doctoral students who have been a constant source of support, inspiration, and friendship. Our shared struggles, late-night discussions, and moments of breakthrough have made this journey not only bearable but truly enjoyable.

To my family and friends, thank you for your unconditional love, understanding, and encouragement. Your belief in me, especially during challenging times over these three years, has been a source of strength and motivation.

This thesis stands as a testament to the collaborative nature of academic pursuit, and I am profoundly grateful for the opportunity to have been part of this enriching experience.

FUNDING: This thesis 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 French government through the France 2030 investment plan managed

by the National Research Agency (ANR), as part of the Initiative of Excellence Université Côte d'Azur under reference number ANR-15-IDEX-01. I am grateful to the Université Côte d'Azur's Center for High-Performance Computing (OPAL infrastructure) for providing resources and support.

Contents

List of Publications

Peer-reviewed international conferences:

[1] **Pierpaolo Goffredo**, Elena Cabrio & Serena Villata, 2024. "Repairing Fallacious Argumentation in Political Debates". In *ACM/SIGAPP Symposium on Applied Computing.* (under review)

[2] **Pierpaolo Goffredo**, Mariana Espinoza, Serena Villata & Elena Cabrio, 2023. "Argument-based Detection and Classification of Fallacies in Political Debates". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Available: https://doi.org/10.18653/v1/2023.emnlp-main.684

[3] **Pierpaolo Goffredo**, Elena Cabrio, Serena Villata, Shohreh Haddadan & Jhonatan Torres Sanchez, 2023. "Disputool 2.0: A modular architecture for multi-layer argumentative analysis of political debates". In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 13, pages. 16431-16433, Washington, DC, USA. Available: https://doi.org/10.1609/aaai.v37i13.27069

[4] **Pierpaolo Goffredo**, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio & Serena Villata, 2022. "Fallacious Argument Classification in Political Debates". In *International Joint Conferences on Artificial Intelligence Organization*, pages 4143-4149, Vienna, Austria. Available: https://doi.org/10.24963/ijcai.2022/575

Peer-reviewed international journals:

[1] **Pierpaolo Goffredo**, Elena Cabrio & Serena Villata, 202X. "Argument extraction and classification on 60 years of U.S. political debates". In *Artificial Intelligence Review.* (under submission)

List of Figures

List of Tables

List of Abbreviations

NLP	**N**atural **L**anguage **P**rocessing
CL	**C**omputational **L**inguistics
AI	**A**rtificial **I**ntelligence
ML	**M**achine **L**earning
ADU	**A**rgumentative **D**iscourse **U**nits
SDI	**S**trategic **D**efense **I**nitiative
SOTU	**S**tate **O**f **T**he **U**nion
BOW	**B**ag **O**f **W**ords
TF-IDF	**T**erm **F**requency-**I**nverse **D**ocument **F**requency
ELMo	**E**mbeddings from **L**anguage **Mo**dels
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
GPT	**G**enerative **P**re-trained **T**ransformer
MLM	**M**asked **L**anguage **M**odeling
NSP	**N**ext **S**entence **P**rediction
NLI	**N**atural **L**anguage **I**nference
LLM	**L**arge **L**anguage **M**odel
IAA	**I**nter-**A**nnotator **A**greement
BIO	**B**eginning, **I**nside, **O**utside
RNN	**R**ecurrent **N**eural **N**etworks
GRU	**G**ated **R**ecurrent **U**nit
CRF	**C**onditional **R**andom **F**ield
DH	**D**igital **Humanities**
NER	**N**amed **E**ntity **R**ecognition
PoS	**P**art **o**f **S**peech
SOTA	**S**tate-**O**f-**T**he-**A**rt

Chapter 1

Introduction

This chapter outlines the motivation behind the present work, emphasizing the need for automatic tools to analyze political debates. It explains why Argument Mining (AM) methods are suitable for addressing this challenge, particularly when combined with frameworks for argument structure analysis. The chapter highlights the potential of AM to enhance understanding of political argumentation and identifies the current research gap in applying comprehensive AM methods to political debates. Finally, it provides an overview of the thesis structure.

1.1 Background & Motivation

The fundamental human capability of argumenting in natural language has ancient roots, tracing back to Socratic dialogues in the Athenian Agora. Throughout history, argumentation evolved through various forms and forums. The printing press revolutionized the spread of ideas, allowing for broader participation in intellectual discourse. The 20th century saw the rise of mass media, with radio and television providing new platforms for public debate. Finally, the digital age has ushered in an unprecedented era of global connectivity, extending argumentation to contemporary discussions on social media and online forums. Throughout this journey, argumentation has assumed a crucial role in various domains, including legal, scientific, educational, medical, and, increasingly in recent years, political spheres [44].

This pervasive nature of argumentation has resulted in a vast amount of argumentative text across various media and platforms. While argumentation scholars have traditionally analyzed these texts manually, the big volume of available data has created a need for more efficient processing methods. Natural Language Processing (NLP) has therefore emerged to address this challenge, enabling the automated analysis of large-scale argumentative corpora. This development has, in turn, sparked the need for specialized methods tailored to argumentative text, leading to the emergence of dedicated Argument Mining (AM) tasks.

Mochales and Moens [117] introduced AM within the context of legal informatics, but it has since expanded to various domains. AM is defined as "the general task of analyzing discourse on the pragmatic level and applying a certain argumentation theory to model and automatically analyze the data at hand" [54]. In this context, argumentation is examined from the perspective of computational linguistics, with the objective of detecting, classifying, and evaluating the quality of argumentative structures within texts. The development of AM methods has led to focus on several challenges, and to the definition of subtasks. Standard AM tasks include identifying argumentative components (premise and claims) and their boundaries in unstructured text, as well as predicting relationships (attack and support) between these components. These tasks are complex and require sophisticated approaches. The methodologies developed in this field aim to aggregate, synthesize, and structure arguments within texts; summarize and reason about argumentation; enable users to search for specific arguments and their justifications; allow for systematic retracing of argumentation; and provide a more solid foundation for decision-making processes and discourse analysis. Given that in-depth manual analysis of texts is time-consuming and requires specialized knowledge, thus proving to be poorly scalable, there is a pressing need to develop automated or machine-assisted approaches for argument extraction and processing.

One particularly rich application area for AM is in the context of political debates as they represent a valuable source of data for argumentative text. In these contexts, candidates employ arguments to justify past actions, present future plans, and critique their opponents' positions. The analysis of argumentation in political discourse primarily requires the reconstruction of argumentative structures formed by speakers. These structures are represented by argumentative components (e.g., premise, claim) and their potential relationships (e.g., attack, support). In the political sphere, AM approaches emerge as valuable not only for analyzing argumentative structures, but also for identifying and categorizing fallacious reasoning. The presence of fallacies is a crucial element in evaluating the effectiveness and validity of presented arguments, highlighting a close correlation between argumentative structure and the presence of logical errors [30]. A fallacy manifests when, despite an apparently correct argumentative structure, the logical connection between its constituent elements is lacking or absent.

The application of AM and related NLP tasks in the political context offers multiple challenges, including automatic information extraction, analysis of argumentative structure, identification of rhetorical patterns, and support for comparative analysis. Furthermore, AM can be effectively integrated with other text analysis modules, further enhancing its utility. For example, interaction with systems for detecting argumentative fallacies can provide a more comprehensive assessment of the quality and validity of presented arguments.

Despite the existence of numerous AM approaches and annotated corpora, few studies have applied argument mining to political texts, and the problem of extracting

complete argumentative structures from such texts has only been partially addressed [61]. This gap represents a significant opportunity for current research in the field of Argument Mining applied to political discourse. The development of such methods could not only enrich the understanding of the political process but also offer new perspectives for researchers, political analysts, and citizens interested in a more objective and structured evaluation of public discourse.

This Ph.D. thesis aims to address the aforementioned research gap by focusing on the case study of American political debates. The central objective is to develop and implement novel algorithms for the detection and extraction of complex argumentative structures within these debates. Furthermore, this research seeks to integrate these extracted structures with additional relevant information, such as the presence and types of argumentative fallacies to better equip both citizens and policymakers to critically evaluate the information they encounter, particularly in the context of political debates.

1.2 Research Questions

To address the identified gaps and limitations in previous studies, the work carried out in this thesis has focused on the following research questions (RQs). These questions guide the investigation throughout this thesis, with each subsequent chapter addressing and providing an answer to one or more of these RQs:

RQ1: *How can Argument Mining tools be designed to effectively support users in exploring and analyzing argumentative structures in political debates?* This question aims to investigate how automated argument mining methods can be integrated into accessible platforms, enabling researchers to effectively explore, visualize, and evaluate argumentative structures in political debates.

RQ2: *How can argumentative structures in political debates be effectively identified and analyzed using computational methods?* This question aims to explore the development of automated methods for recognizing and examining the argumentative components and relationships within political debates. It addresses the fundamental challenge of translating the complex and often implicit argumentation patterns found in political discourse into structured, machine-readable formats.

RQ3: *How can computational approaches be developed to classify and detect fallacious arguments in political debates?* This question can be further subdivided into two more specific sub-questions:

- *How can transformer-based models be effectively employed to automatically detect and classify different categories of fallacious arguments in political debates?* This sub-question addresses the core methodological approach of using advanced natural language

processing methods (specifically, transformer-based models) to identify and categorize fallacies in political debates. It combines the automatic classification of snippets in political debates already classified as fallacy, and it further analyzes the approach to identify and classify at the same time in a broader context of the text.

- *To what extent does the incorporation of argumentative structure and linguistic features enhance the accuracy of fallacy detection and classification models in political debates?* This sub-question explores the impact of including additional contextual information in the analysis. It combines argument components and relations, and the inclusion of textual features, as PoS tags.

RQ4: How can fallacious arguments in political debates be effectively repaired into valid, non-fallacious arguments, and to what extent can Large Language Models be employed in this process? This research question aims to develop methods for transforming misleading arguments in political discourse into logically sound alternatives without changing the meaning. It explores methods to analyze argumentative structures, identify fallacies, and generate improved versions using advanced LLMs, and the human effort to evaluate the LLMs' answers.

1.3 Contributions

Answering to the RQ mentioned above, the main contributions of the thesis are as follows:

Contribution 1 — Updated Argument Mining Tool for Political Debates Analysis with New Dataset, Views, and Argument Mining Pipeline This contribution presents a significant advancement in argument mining, directly addressing RQ1 by providing a comprehensive analysis tool for the field. This upgraded version integrates improved argumentative component identification, relationship mapping, Named Entity Recognition, and graph overviews, offering a holistic approach to text analysis based on the first version of the ElecDeb60to20 dataset [51]. The tool's improvements, including an upgraded core model and expanded analytical capabilities for detecting argument fallacies, demonstrate its commitment to providing deeper insights into complex argumentative structures.

Related Publications:

1. **Pierpaolo Goffredo**, Elena Cabrio, Serena Villata, Shohreh Haddadan & Jhonatan Torres Sanchez, 2023. "Disputool 2.0: A modular architecture for multi-layer argumentative analysis of political debates". In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 13, pages. 16431-16433, Washington, DC, USA.

Contribution 2 — Domain-Specific Approach to Apply Argument Mining for Political Debate Analysis It addresses RQ2 by proposing a domain-specific definition of Argument Mining tasks. The approach presents a comprehensive two-stage Argument Mining pipeline integrating both component detection and relation classification to analyze political debates. The first stage employs advanced neural networks to identify argument components within text, while the second stage predicts relationships between these components. This pipeline demonstrates practical effectiveness, surpassing standard baselines in both component detection and relation prediction.

Related Publications:

1. **Pierpaolo Goffredo**, Elena Cabrio & Serena Villata, 202X. "Argument extraction and classification on 60 years of U.S. political debates". In *Artificial Intelligence Review.* (under submission)

Contribution 3 — Updated ElecDeb60to20 Dataset with Argumentative and Fallacious Annotations To address RQ3, the ElecDeb60to16 [63] dataset was significantly expanded and updated. This updated dataset, named ElecDeb60to20, serves as a crucial resource for training and evaluating classifiers in supervised argument mining approaches. The expansion includes the 2020 U.S. presidential election between Biden and Trump, adding new argumentative components, relations, and fallacious arguments to the corpus. This update ensures the dataset remains current and representative of contemporary discourse in the field. The expanded dataset, now one of the largest annotated collections in its domain, offers researchers a comprehensive resource for exploring various aspects of argumentation, from basic structures to complex fallacious reasoning patterns. It captures the evolution of discourse over an extended period, making it valuable for developing and evaluating novel methods in argument extraction, classification, and analysis.

Related Publications:

1. **Pierpaolo Goffredo**, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio & Serena Villata, 2022. "Fallacious Argument Classification in Political Debates". In *International Joint Conferences on Artificial Intelligence Organization*, pages 4143-4149, Vienna, Austria.

Contribution 4 — Modelling Detection and Classification of Fallacies in Political Debates To address RQ3, advanced machine learning models are developed and evaluated for fallacy detection and classification in argumentative texts. The initial approach enhanced existing language models in multi-class classification by incorporating argumentative feature labels, significantly improving performance in multi-class classification. Building on this, a novel model was created to solve the classification and detection task, integrating various linguistic features such as argument components, relations, and Part-of-Speech tags. This model achieves improved performance in both detecting fallacy boundaries and classifying them, outperforming existing approaches.

Related Publications:

1. **Pierpaolo Goffredo**, Mariana Espinoza, Serena Villata & Elena Cabrio, 2023. "Argument-based Detection and Classification of Fallacies in Political Debates". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore.

Contribution 5 — Creation of a New Dataset of Annotated Repaired Fallacies To address RQ4, FallacyFix: A Repaired Fallacies Dataset has been created, addressing the challenge of repairing fallacious reasoning in argumentative texts in political debates. Built upon the ElecDeb60to20-fallacy dataset, this new resource focuses on systematically repairing fallacious arguments, particularly those involving common fallacy types in political debates. This dataset represents a significant contribution to the field of Argument Mining, offering researchers a unique resource for developing and evaluating methods to analyze and repair fallacious reasoning in various domains where argumentation is crucial.

Related Publications:

1. **Pierpaolo Goffredo**, Elena Cabrio & Serena Villata, 2024. "Repairing Fallacious Argumentation in Political Debates". In *ACM/SIGAPP Symposium on Applied Computing.* (under review)

1.4 Structure

The thesis is organized as follows:

Chapter 2 explores Argument Mining (AM) and its application to political discourse analysis. It covers core AM principles, focusing on fallacies in political rhetoric and their key categories. The chapter examines recent AM developments, including classification techniques, detection methods, and language model applications. It contextualizes the thesis's contributions within current AM research, highlighting its significance in fallacy studies and political debate analysis.

Chapter 3 presents ElecDeb60to20, a dataset of U.S. presidential debate transcripts from 1960 to 2020. It includes annotations for argument components, relations, fallacies, and fallacy repairs. Initially covering debates up to 2016, it was expanded to include the 2020 Biden-Trump debate. This dataset underpins all thesis experiments, offering insights into political discourse evolution over six decades.

Chapter 4 introduces an Argument Mining pipeline for political debates, focusing on argument component detection and relation prediction. It compares methods from Support Vector Machines to fine-tuned Transformers with RNNs for component detection, and innovatively treats relation classification as sequence classification. The

chapter analyzes performance results and errors, highlighting challenges in capturing argumentative structures in political discourse, especially in classifying attack relationships and processing concise statements.

Chapter 5 presents DispuTool 1.0 and 2.0, a tool for automated argumentative analysis of political debates. This tool identifies argument components and relationships, visualizing debates as interactive graphs. DispuTool features entity recognition, fallacy visualization, and analytical graphics. The chapter demonstrates its practical application in U.S. Presidential Debates.

Chapter 6 focuses on multi-class classification and detection of argumentative fallacies in political debates. It explores transformer-based models for fallacy classification using ElecDeb60to16 and addresses fallacy detection using ElecDeb60to20. The chapter highlights the integration of argumentative and textual features to improve model performance.

Chapter 7 examines Argument Mining in political debates, focusing on fallacy repair. It introduces FallacyFix, a dataset of human-repaired fallacious arguments, to evaluate large language models (LLMs). The chapter assesses LLMs' ability to repair and classify various fallacies in debates using a modular prompt structure. Evaluation involves automated metrics and human assessment, comparing model-generated repairs to gold standards.

Chapter 8 concludes the thesis by summarizing its key contributions, addressing open questions, and proposing future research directions. It outlines potential applications of the developed methodologies and discusses plans for further improvements, enclosing the thesis's achievements while setting the stage for continued advancements in political discourse analysis.

Chapter 2

Background

This chapter shows the theoretical foundations and methodological advancements in Natural Language Processing (NLP) that underpin the analysis of political discourse. The discussion starts with an examination of natural language representation approaches, tracing their evolution from rule-based systems to sophisticated machine learning paradigms. Subsequently, the chapter explores the field of Argument Mining, its specific application within the domain of political debates, and the critical area of fallacy analysis.

The field of Natural Language Processing has evolved significantly since its inception in the mid-20th century, transitioning from rule-based symbolic systems to approaches rooted in statistics and Machine Learning. This evolution has paved the way for more sophisticated analysis of human communication, including the emerging field of Argument Mining.

2.1 Natural Language Representation

Initially dominated by rule-based symbolic systems, the NLP field has since seen approaches rooted in statistics and Machine Learning gain prominent position [82, 74]. Machine Learning (ML) aims to develop mathematical models trained on sample data to make predictions about new, unseen information. A key challenge in this process is the quantification of data for use by Machine Learning models, which is particularly crucial in NLP, as human communication is not inherently numerical. This process is fundamental for a range of tasks, including machine translation, natural language generation, and text classification [82, 75]. The conversion of language into numerical form presents numerous challenges. Languages differ not only in vocabulary but also in fundamental structure [27]. As a result, representation models effective for one language may not be suitable for others. Furthermore, natural language is characterized by high levels of ambiguity and context-dependence. Words can represent multiple concepts depending on their context, and the implications and meaning of a sentence are heavily influenced by both context and the speaker's intent [131]. The fact that even human communication can lead to misunderstandings underscores the complexity of

language comprehension. Despite significant advancements, achieving a complete understanding of natural language remains an unresolved challenge in NLP. The numerical representation of language while preserving its subtleties and context-dependency continues to be an active research area. Nevertheless, the past decade has witnessed substantial progress in natural language representation. Subsequent discussions typically highlight the primary approaches developed to address this ongoing challenge in the field of NLP.

2.1.1 Context-free Representation

BOW. The Bag of Words (BOW) model is a fundamental approach in Natural Language Processing that represents text by focusing on word frequency within a document, avoiding grammatical structure and word order to capture the essence of content through its keywords. Originally developed for information retrieval tasks like document search, the BOW model transforms each text into a numerical vector, where each dimension corresponds to a unique word from the overall vocabulary, and the value represents that word's frequency in the text. This approach results in high-dimensional, sparse vectors, as most words in the vocabulary won't appear in any given document. To enhance efficiency, preprocessing methods like removing stop words or lemmatization can be employed to reduce vocabulary size, though it often remains substantial for large corpora. While the BOW model excels in simplicity and effectiveness for certain tasks, it has limitations: by disregarding word order and relationships, it sacrifices semantic nuance, cannot differentiate between synonyms, struggles with word sense disambiguation, and fails to capture contextual meaning arising from word combinations. Despite these constraints, BOW serves as a foundational technique in text analysis, providing a straightforward yet powerful method for quantifying textual data and enabling various downstream Natural Language Processing tasks.

TF-IDF. Term Frequency-Inverse Document Frequency (TF-IDF) is an advanced text representation method that builds upon the Bag of Words model by introducing a weighting scheme based on word rarity, resulting in more meaningful vector representations. This approach assigns higher weights to words that appear frequently in a specific document but are rare across the entire corpus, thus emphasizing distinctive terms that characterize a document's content. Conversely, common words like "the", which appear ubiquitously across documents, receive lower weights, reducing their impact on the overall representation. Originally developed to enhance search engine result relevance by highlighting key terms in pertinent documents, TF-IDF has become a widely used technique in various Natural Language Processing tasks. However, like its predecessor, the Bag of Words model, TF-IDF still lacks semantic understanding and cannot discern nuanced word meanings or contextual relationships between terms. Despite this limitation, TF-IDF represents a significant improvement in capturing document-specific information and remains a valuable tool in text analysis and information retrieval applications.

N-grams. Understanding the meaning of words often depends on their surrounding context. Unlike earlier methods such as bag-of-words, which treated words in isolation, N-grams capture this crucial contextual information by considering sequences of adjacent words. This technique emerged from the need to properly interpret multi-word expressions and idioms, where meaning arises from specific word combinations. N-grams come in various sizes, with *bigrams* (two-word sequences) and *trigrams* (three-word sequences) being the most common. These shorter N-grams can effectively handle many phrases and negations, while longer N-grams may be used to represent the broader context. However, increasing N also exponentially increases the number of possible combinations, potentially leading to data sparsity issues. Beyond improving semantic understanding, N-grams serve as simple statistical language models, predicting word probabilities based on their context. They can be enhanced using familiar text processing methods like stop word removal, lemmatization, and TF-IDF weighting to focus on the most informative word sequences. Despite their usefulness in capturing local context, N-grams have limitations. They are unable to model long-range semantic dependencies/relationships between words separated by larger distances in the text. These dependencies are often crucial for tasks like pronoun resolution or disambiguating word meanings based on broader context. While variations like skip-grams attempt to address this weakness, modeling complex semantic relationships across an entire document remains challenging for N-gram-based approaches.

Word Embedding. While document-level encoding is valuable, representing individual words effectively is equally crucial. The simplest approach, one-hot encoding, creates sparse vectors with a single '1' at the word's index and '0's elsewhere. However, this method produces high-dimensional, sparse vectors incorrect for neural networks. Word embeddings address these limitations by representing words as dense, lower-dimensional vectors (typically 300 dimensions) that capture semantic relationships. These embeddings are learned by observing word usage across large text corpora, creating a semantic vector space where similar words cluster together and relationships between words are preserved. In this vector space, remarkable properties emerge. For instance, vector arithmetic can model analogies: "King" - "Man" + "Woman" results in a vector close to "Queen". This capability extends to various semantic relationships and even handles polysemous words effectively. As Neelakantan et al. [123] explains, "In moderately high-dimensional spaces a vector can be relatively *close* to multiple regions at a time". However, word embeddings are not without challenges. They can inherit biases present in training data, potentially altering word representations (e.g., "apple" might lean towards its corporate meaning if trained on business news). Additionally, words absent from the training corpus lack embeddings, leading to out-of-vocabulary issues. Pre-trained embeddings, such as those from Word2Vec [115] or GloVe [129] models, offer advantages in machine learning applications. Trained on diverse, general corpora, they provide broadly applicable representations that transfer well across tasks, especially valuable when task-specific data is limited. However, for

specialized domains like biomedicine, where vocabulary and semantics differ significantly from general language, these pre-trained embeddings may underperform. In such cases, domain-specific embeddings can be learned from relevant texts, capturing nuanced meanings and terminology. This approach typically improves performance but requires sufficient in-domain data. The choice between using *pre-trained embeddings*, *fine-tuning* them on domain-specific data, or *training* embeddings from scratch depends on the available data and the specific requirements of the task. Throughout this thesis, various types of word embeddings were explored as input representations for neural networks and other machine learning models, balancing the trade-offs between general and domain-specific representations.

2.1.2 Context-dependent Representation

Modern, up-to-date text representation methods employ *contextualized* embeddings to overcome the limitations of static word embeddings. While static embeddings capture general semantic information, they fail to account for the surrounding context and assign a single, fixed vector to each word. This rigid representation doesn't adapt to context, leading to challenges with polysemy — words with multiple meanings are condensed into a single vector, leaving downstream models to interpret the intended sense. The importance of context in determining word semantics becomes evident in tasks like sentiment analysis. For instance, accurately classifying the sentiment of "The bank closed my account" requires understanding whether "bank" refers to a financial institution or a riverbank based on the surrounding words. Static embeddings struggle to differentiate between these meanings, highlighting a significant shortcoming.

Contextualized embeddings, on the other hand, are *dynamic*, generating context-aware representations tailored to each specific instance. These embeddings can encode the word "bank" differently depending on the surrounding text, providing the necessary disambiguation for accurate sentiment analysis [143, 121]. Rather than assigning a single vector per word, contextualized embeddings encode words differently based on their linguistic context in each occurrence. This approach more accurately reflects the variability in word usage and meaning across different contexts. By producing representations that adapt to the specific context in which a word appears, contextualized embeddings offer a more nuanced and accurate way to capture the intended semantics of words in the text. This context-sensitivity is crucial for tasks that require fine-grained understanding of word meanings and their variations across different uses.

ELMo. Peters et al. [130] introduced a groundbreaking approach to generate contextualized word representations called Embeddings from Language Models (ELMo). In contrast to traditional static embeddings, ELMo constructs dynamic word representations that adapt to the surrounding sentence context. The core of ELMo lies in its bidirectional language model (biLM) architecture, which consists of two distinct but interconnected LSTM-based neural networks [70]. The forward LSTM processes the sentence sequentially from left to right, while the backward LSTM operates in the

opposite direction, from right to left. Each LSTM layer captures contextual information from its respective direction, enabling the model to learn long-range dependencies within the input sequence. The LSTMs require pre-training on a large text corpus, learning to predict the next word based on the preceding context. To create the final ELMo word vectors, the internal representations from both the forward and backward LSTMs are concatenated along the depth dimension. This concatenation allows ELMo to incorporate both past and future context when generating each word embedding. Furthermore, by combining representations from multiple LSTM layers, ELMo captures a range of linguistic information, from low-level syntactic features to higher-level semantic relationships. As a result, ELMo produces rich, context-dependent word vectors that dynamically adapt to the specific sentence, marking a significant advancement over previous static embedding approaches.

Transformer. Vaswani et al. [156] introduced the Transformer model, which revolutionized the architectural landscape of Natural Language Processing tasks. The Transformer model breaks away from its predecessors by introducing the concept of self-attention, also known as *scaled dot-product attention*. This mechanism enables the model to assess the relevance of each word in a sentence when encoding a specific word, allowing it to effectively capture the sentence's context. Unlike recurrent models like LSTMs, which process sentences sequentially, Transformers process all words in the sentence concurrently. This parallel processing approach facilitates more efficient computation and better handling of long-range dependencies. The Transformer model consists of two primary components: *(1)* an encoder that processes the input text and *(2)* a decoder that generates the output text word by word. Each component is composed of multiple layers, including self-attention mechanisms and feed-forward neural networks, enabling the model to discover intricate patterns within the data. Notably, the decoder generates each word by considering the encoder's output and its own previously generated words. This groundbreaking architecture has become the foundation for numerous subsequent models, such as BERT and GPT, which have further advanced the field of NLP. Figure 2.1 represents the Transformer model architecture.

GPT. Radford et al. [132] introduced the Generative Pretrained Transformer (GPT), a groundbreaking contextual embedding technique that has had a significant impact on the field of Natural Language Processing. GPT leverages the decoder component of the Transformer model and undergoes pre-training using a unidirectional language modeling objective. This pre-training approach enables GPT to generate text with remarkable proficiency, making it particularly well-suited for tasks that involve text generation. However, due to its unidirectional nature, GPT can only consider context from one direction when generating embeddings. While this unidirectional approach suffices for many applications, it may constrain the model's ability to fully grasp the subtle semantic nuances present in language. This limitation is subsequently addressed by the development of the BERT model, which incorporates bidirectional context for a more comprehensive understanding of language.

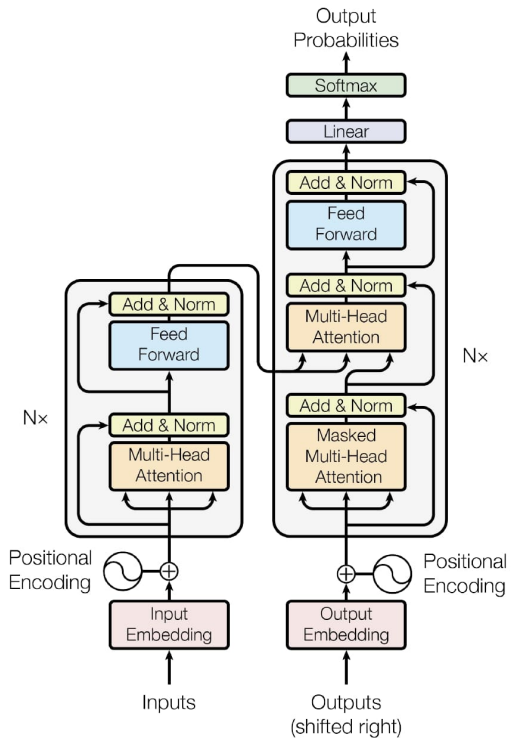


Figure 2.1: Transformer model architecture. Figure drawn from Vaswani et al. [156].

BERT. Devlin et al. [36] revolutionized the field of contextualized embeddings with the introduction of the Bidirectional Encoder Representations from Transformers (BERT) architecture. Unlike previous models such as ELMo or unidirectional models like GPT, BERT employs the Transformer encoder bidirectionally. This unique feature enables BERT to comprehend the context of a word by considering both the preceding and following words, making it highly effective in disambiguating word meanings based on their surrounding context. BERT undergoes pre-training using two primary objectives: *(1)*Masked Language Modeling (MLM) and *(2)*Next Sentence Prediction (NSP). The MLM objective, inspired by the Cloze task [149], involves randomly masking words in the input and tasking the model with predicting the original masked words. Unlike the unidirectional next-word prediction task, the MLM objective allows BERT to consider both the left and right context simultaneously. The second pre-training task, NSP, assesses the model's ability to understand the relationship between two sentences, which is particularly valuable for downstream NLP tasks that require an understanding of sentence relationships, such as Question Answering. In the NSP task, the model is presented with two sentences and must determine whether the second sentence logically follows the first in the original text. Following pre-training, BERT can be fine-tuned for specific downstream tasks. Fine-tuning involves continuing the training process on a specific task (such as sentiment analysis or question answering) using a smaller, task-specific dataset. This process adjusts the pre-learned representations to better align with the specific task, leveraging the broad language understanding acquired during

pre-training while adapting to task-specific patterns. Fine-tuning is relatively cost-effective compared to pre-training, making BERT a versatile and efficient model for a wide range of NLP tasks. The remarkable versatility and effectiveness of the BERT model have inspired the development of numerous specialized models that build upon the original architecture or adapt it to specific domains. These models harness the power of BERT while tailoring its capabilities to better suit specific tasks or types of data. For example, domain-specific models such as SciBERT [12], BioBERT [93], PubMedBERT [52], and POLITICS [104] are trained on scientific, biological, medical, and political text respectively. These models excel at capturing domain-specific language and semantics, significantly improving performance on tasks within these fields. In addition to domain-specific models, architectural advancements have also been made to enhance the base BERT model. For instance, RoBERTa [103] refines the BERT training procedure to improve its performance by training with larger mini-batches, utilizing more data, and eliminating the next sentence prediction task. ALBERT [89], on the other hand, reduces the model size of BERT while maintaining comparable performance, making it more efficient. These models exemplify the ongoing evolution of BERT and its lasting impact on the field of Natural Language Processing.

SBERT. Sentence-BERT (SBERT) [134] represents a groundbreaking advancement in the field of contextualized models, transcending the limitations of word-level embeddings to capture the semantics of entire sentences. While models like BERT generate embeddings for individual words, SBERT takes a leap forward by creating embeddings that encapsulate the meaning of complete sentences. This is achieved by adapting the BERT architecture to process pairs of sentences and training it on Natural Language Inference (NLI) tasks. In these tasks, the model learns to classify sentence pairs into categories such as *entailment*, *contradiction*, and *neutral*, effectively acquiring an understanding of the semantic relationships between sentences. Through this training process, SBERT learns to encode not only the context and semantic relationships between the words within each sentence but also the relationships between different sentences. The resulting sentence embeddings capture the overall semantic content of each sentence, enabling efficient and meaningful comparisons of semantic similarity between sentences.

This shift in focus from understanding individual words to comprehending the larger units of meaning in language provides a more comprehensive and holistic perspective on textual data. By capturing sentence-level semantics, SBERT opens up new possibilities for applications that require an understanding of the overall meaning of sentences. Tasks such as semantic search, text clustering, and paraphrase detection can greatly benefit from the rich semantic information encoded in SBERT embeddings. This advancement in sentence representation marks a significant step forward in the field of Natural Language Processing, enabling more sophisticated and nuanced analysis of textual data at a higher level of abstraction.

Large Language Models The development of models like BERT and GPT has paved the way for the growth of large language models (LLMs) , which have taken the field of Natural Language Processing to new heights. LLMs are a class of deep learning models that are trained on massive amounts of textual data, often ranging from billions to trillions of words, enabling them to capture the intricacies and nuances of human language at an unprecedented scale. LLMs build upon the architectures and training methods introduced by models like BERT and GPT, leveraging the power of the Transformer architecture [156] and self-supervised learning objectives. However, what sets LLMs apart is their sheer scale and the breadth of their knowledge. By training on vast and diverse datasets, LLMs develop a deep understanding of language patterns, semantics, and world knowledge, allowing them to perform a wide range of NLP tasks with remarkable proficiency. One of the key advantages of LLMs is their ability to generalize and adapt to various tasks without extensive task-specific training. Through a process called pre-training, LLMs learn to capture the underlying structure and meaning of language by predicting missing words or generating text based on the given context. This pre-training phase allows LLMs to acquire a broad understanding of language that can be transferred to specific NLP tasks with minimal fine-tuning. The emergence of LLMs has led to significant breakthroughs in various areas of NLP, such as language generation, question answering, and text summarization. Models like GPT-3 [19], one of the first prominent LLMs, have demonstrated the ability to generate human-like text, engage in coherent conversations, and even perform creative writing tasks. The success of GPT-3 has sparked a wave of research and development in the field of LLMs, with researchers exploring new architectures, training methods, and applications. However, the development of LLMs also presents challenges and considerations. The potential for biases in the training data and the ethical implications of using LLMs in real-world applications need to be carefully addressed to ensure responsible and fair use of these powerful models.

As natural language representation methods have advanced, they have enabled more complex NLP tasks, including the analysis of argumentative structures in text. This progression has led to the development of Argument Mining, a specialized subfield that leverages these representation methods to identify and analyze arguments in natural language.

2.2 Argument Mining

Argument Mining, a specialized subfield of Natural Language Processing, automatically integrates AI methodologies to identify and analyze argumentative structures in the text. This interdisciplinary domain leverages NLP methods with computational argumentation and cognitive science insights. This domain focuses on analyzing the process of constructing, comparing, evaluating, and analyzing arguments, with the ultimate goal of determining their validity [11, 14]. Its application extends to resolving various theoretical and practical issues, including the explanation and justification

of decisions, as well as reasoning in contexts characterized by inconsistent or incomplete information. In general terms, an argument can be defined as a set of premises or claims that, through a logical process, lead to a conclusion. The primary purpose of argumentation is to influence the degree of acceptability of certain statements, either by reinforcing them or by challenging them through new arguments. However, the practical application of argumentation in decision-making processes requires structured inputs. In reality, argumentative texts often present themselves in an unstructured form, lacking explicitly identifiable argumentative components. This situation has necessitated the development of computational methods for the automatic extraction of structured arguments from raw texts. In response to this need, one of the most significant advancements in the field of artificial argumentation [9] is represented by Argument Mining (AM) [101, 90].

This is a discipline that tackles the complex task of automatically extracting and analyzing argumentative structures from natural language texts. This field focuses on developing sophisticated systems capable of identifying key components of arguments, such as claims and premises, as well as discerning the intricate relationships between these elements[101, 112]. It finds application in various fields, ranging from evaluating the persuasive effectiveness of essays to interpreting legal reasoning, from analyzing clinical studies to combating disinformation and examining political debates [101, 53, 109, 33, 21]. Among these pioneering works, argumentative zoning [151] introduced the classification of sentences based on their rhetorical role within scientific publications, paving the way for subsequent AM methodologies [90]. Additionally, other studies focused on identifying argumentative components in legal texts [118, 117]. However, these initial approaches were constrained by the limitations of NLP methods available at the time. Advancements in Natural Language Processing computational methods have enabled more complex tasks like Argumentation Mining, leading to a significant increase in interest in this field [21]. Argument Mining requires a deep understanding of natural language and exhibits close connections with linguistic inference. It is not coincidental that early methods in this field have taken inspiration from textual entailment [20, 17]. The current maturity achieved by NLP has opened new perspectives, enabling researchers to develop increasingly effective and innovative Argumentation Mining methodologies.

Argument Mining is commonly conceptualized as a two-phase process, articulated in two fundamental and interconnected operations:

1. *Identification of Argumentative Components*: The initial phase focuses on identifying and delimiting the constituent elements of argumentation within a textual corpus. These elements include claims and premises. This identification process can be further decomposed into two distinct but complementary sub-operations:

   (a) **Detection**: identifying the presence of argumentative components in the text.

(b) **Segmentation**: precisely demarcating the boundaries of each identified component ("Argumentative Discourse Units" (ADU) [128], represent the minimal units of analysis in the context of AM).

2. *Analysis of Argumentative Relations:* The second phase focuses on mapping the interactions between the previously identified argumentative components. This analysis can be done using two different methods. In one approach, the goal is to pair the identified argumentative structures with predefined schemas, known as "argumentative" or "rhetorical" schemas, which represent recurring patterns of reasoning [164]. Alternatively, one can apply for a more flexible analysis, where relationships between argumentative components are evaluated independently, without reference to pre-established schemas. This process aims to decipher the nature of the logical connections that bind the various elements, classifying them into categories such as:

   (a) **Support**: when one element strengthens or corroborates another.

   (b) **Attack**: when one element contrasts with or refutes another.

The final phase reconstructs the overall argumentative structure by analyzing relationships between components. This can be done either by matching identified structures to predefined argumentative schemas or through a more flexible analysis of component relationships without reference to established patterns. The integration of these approaches enables the construction of comprehensive graphical representations of the argumentative structure, offering a holistic and systematic view of the analyzed discourse [45, 119]. The ultimate goal is to comprehensively map out the argumentative content and structure of the text, providing a clear and detailed understanding of the discourse's logical flow and persuasive elements.

Figure 2.2 shows an example of an Argumentation Mining pipeline in the context of political debates.

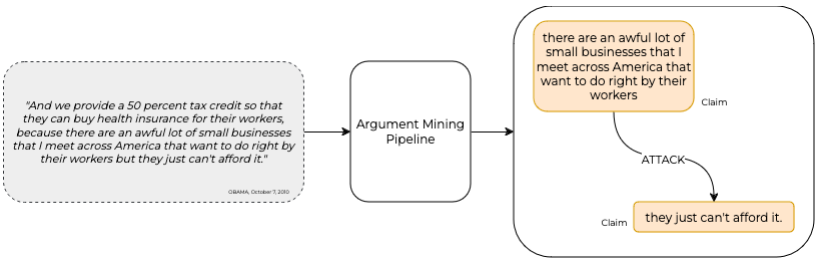


Figure 2.2: Example of Argumentative Mining pipeline.

The evolution of the field has also led to the emergence of more sophisticated and nuanced subtasks that go beyond the straightforward identification of argumentative components and structures. These supplementary tasks aim to enrich the analysis with additional informative features, particularly useful in concrete application scenarios. These include:

- **Argument clustering** groups similar arguments together, which is particularly useful in analyzing large-scale debates or discussions. For instance, in a public consultation on a new city policy, this technique could help policymakers quickly identify the main themes of public concern by grouping similar arguments from thousands of citizen submissions [135].
- **Evaluation of argumentative relevance** assesses how pertinent an argument is to the topic at hand. It's particularly crucial in scenarios involving large-scale argument analysis, such as in advanced search engines or debate platforms. The goal is to develop methods that can automatically assess and rank arguments based on their relevance to a given topic or query. These methods often involve analyzing the relationships between different arguments, considering factors like how frequently an argument is referenced or how central it is to the overall discussion. By evaluating argumentative relevance, systems can prioritize the most significant points in a debate, helping users navigate complex argumentative landscapes more effectively [163].
- **Analysis of argument quality** evaluates the strength and persuasiveness of arguments. In academic settings, this could be used to automatically grade student essays, providing feedback on the robustness of their argumentative structures and helping students improve their critical thinking and writing skills [154, 47, 108].
- **Identification of rhetorical figures** involves recognizing and analyzing rhetorical devices used in argumentative discourse. Rhetorical figures, such as repetition patterns or contrasting structures, can serve as indicators of argumentative content and help reveal the underlying reasoning structure of texts. By automatically identifying these figures, Argument Mining methods can potentially improve in performance across various domains, from political speeches to academic writing. This approach not only aids in argument analysis but also offers an opportunity to empirically test theoretical claims about the role of rhetorical devices in argumentation [92].
- **Detailed classification of evidence types** focuses on identifying and categorizing various evidence types in medical literature. For instance, it supports evidence-based medicine by enabling automated detection of claims and evidence from unstructured medical texts. This technique can significantly streamline medical literature reviews, aiding healthcare professionals in making evidence-based decisions efficiently [109].
- **Fallacy identification and classification** involves detecting and categorizing fallacies. In combating online misinformation, this technique could be used to automatically flag posts containing common fallacies, helping social media platforms and users identify potentially misleading content and promote more rational discourse [81, 157, 5, 161, 139].

While Argument Mining has broad applications across various domains, its relevance to political discourse is particularly significant. The analysis of political debates presents unique challenges and opportunities for Argument Mining techniques, as these debates often contain complex argumentative structures and rhetorical strategies.

2.3 Argument Mining on Political Debates

This section presents a critical review of key developments in AM as applied to political discourse, with a focus on methodologies for detecting argument components, identifying argument relations, and analyzing rhetorical strategies. Table 2.1 provides a comprehensive overview of existing annotated resources in this domain.

Dataset	Resource	Agreement	Size
Chakrabarty et al. [22]	Change My View subreddit	Kripendorff's $\alpha = 0.61$ for relation/no relation, $\alpha = 0.63$ for relation types	2,756 sentences
Naderi and Hirst [120]	Canadian parliamentary debates	Weighted $\kappa = 0.54$ for stance, weighted $\kappa = 0.46$ for frames (first), weighted $\kappa = 0.70$ for frames (second)	121 statements
EtHanThatcher3 [38]	UK parliamentary debates	Cohen's $\kappa = 0.67$ ESE/no ESE, $\kappa = 0.95$ for Support/Attack, $\kappa = 1$ for source person, $\kappa = 0.84$ for target person	90,991 words, 638 ESEs
US2016 [158]	U.S. presidential debates, reddit	Cohen's $\kappa = 0.610$ for IAT, CASS $\kappa = 0.752$ for IAT	8937 locutions, 12,965 illoc.
Menini et al. [112]	1960 U.S. presidential speeches	Fleiss' $\kappa = 0.63$ for relation types	1,462 arg. pairs

Table 2.1: Existing annotated datasets for argument mining in political debates.

2.3.1 Component Detection and Relation Prediction

Argument mining in political discourse has been a focus of several seminal studies, each contributing to the field's advancement. Naderi and Hirst [120] made significant strides by annotating Canadian parliamentary debates on same-sex marriage (2005-2006). Their work, centered on stance classification and frame identification, utilized distributed representations of words and sentences along with linguistic features, achieving a peak accuracy of 72.4% in frame prediction. Building on this, Menini et al. [112] developed a more nuanced dataset from the 1960 U.S. presidential election campaign. Their corpus, comprising 1,462 argument pairs, served as the basis for a sophisticated relation classification system. Using a multi-class feature-rich Support Vector Machine, they achieved F-scores of 0.71 for related/unrelated classification and 0.65 for support/attack classification, marking a significant advancement in the field. Expanding the scope of argument mining in political contexts, Duthie, Budzynska, and Reed [39] pioneered the automatic extraction of ethos arguments from UK Parliamentary debates. Their innovative approach, using the Argument Interchange Format (AIF),

demonstrated the feasibility of automated ethos analysis in political discourse, achieving an F-score of 0.60. A more comprehensive approach was taken by Visser et al. [158], who developed an extensively annotated corpus from the 2016 U.S. presidential election debates and associated Reddit discussions. This multi-faceted approach, grounded in Inference Anchoring Theory, represents a significant step toward understanding the complex dynamics of modern political argumentation. [137] evaluated transformer-based models for argument relation detection across domains. Using the US2016 and Moral Maze corpora, they found RoBERTa-large performed best, achieving macro F1-scores of 0.70 in-domain and 0.61 cross-domain. This outperformed previous SOTA results, even with a more complex relation scheme. While not exclusively focused on political debates, other studies have produced results with significant implications for political argument mining. The TARGER framework [24], designed for argument tagging in diverse text types, achieved 64.54% accuracy in argument extraction when trained on multiple datasets, demonstrating the potential for transfer learning across domains. In the realm of social media analysis, Addawood and Bashir [2] achieved an impressive 89% F1-Score in classifying argumentative text using Support Vector Machines with a comprehensive feature set. This work highlights the potential for adapting argument mining methods to the more informal and dynamic context of social media political discourse. Moving beyond politics, recent studies have applied argument mining methods to other domains. [53] presents a novel approach to argument mining in European Court of Human Rights (ECHR) decisions. By developing a legally informed annotation scheme and creating a corpus of 373 annotated ECHR decisions, they trained models that achieved macro F1 scores of 43.13 for argument type detection and 91.36 for agent detection, addressing the gap between legal experts' analysis and NLP approaches. In the realm of social media, [141] examines the state of argument mining on Twitter. This comprehensive survey covers corpus annotation, argument detection, and stance detection, reporting F1 scores ranging from 0.78 to 0.89 for argument detection tasks. The paper also explores the integration of stance detection with argument mining, highlighting the unique challenges posed by the Twitter platform. Chen et al. [23] investigated LLMs in computational argumentation tasks, introducing "counter speech generation." Their results showed that LLMs outperformed baselines on some datasets, achieving high BERTScores (up to 0.92 for GPT-3.5), but faced challenges with strict metrics like ROUGE (scores as low as 0.14 for certain tasks). Finally, in the healthcare domain, [110] presents a transformer-based argument mining pipeline for Randomized Controlled Trial (RCT) abstracts. Their approach, which uses domain-specific BERT models combined with GRU and CRF layers, achieved impressive results with macro F1-scores of 0.87-0.91 for component detection and 0.62-0.69 for relation prediction.

This corresponds to the approach used in Chapter 4 where I adapted the architecture for the main topic of this thesis, the political debates, using the ElecDeb60to20 dataset.

2.3.2 AM Tools

Recent advancements in argument mining have led to the development of several online tools, enhancing accessibility for both researchers and the public. [91] introduced an online annotation assistant that facilitates the identification of argument schemes, crucial for understanding reasoning patterns. MARGOT [102] offers automated extraction of argumentative content from unstructured text via a user-friendly web interface. [145] developed ArgumenText, a sophisticated system for retrieving and analyzing arguments from diverse web sources, enabling comprehensive argument analysis. Complementing these, [162] created args.me, an innovative argument search engine that allows users to explore and compare arguments on various topics across the web.

Unlike previous tools, DISPUTool 2.0 [49] offers a comprehensive suite of argumentative analysis features for political debates, including the novel ability to detect argument components and relations, and exploring the presidential debates from different argumentative points of view, making it a more versatile and advanced solution for researchers and analysts in the field of computational argumentation, as detailed in Chapter 5.

The application of Argument Mining to political debates not only enhances our understanding of political discourse but also reveals the prevalence of fallacious reasoning in these contexts. This insight has led to increased interest in fallacy analysis.

2.4 Fallacy Analysis

The field of Natural Language Processing has experienced significant growth in the detection and classification of fallacies, misinformation, and propaganda [33]. This developing area of research has advanced our understanding of argumentative structures and computational approaches to analyze them.

Standard dictionaries, such as the Oxford English Dictionary, define fallacies as both *invalid arguments* and *faulty reasoning*. This complexity is reflected in how different disciplines approach fallacies: logic focuses on formal invalidity, cognitive science examines biased reasoning, and communication science studies the deceptive and persuasive nature of fallacious discourse [97]. The understanding of fallacies has evolved significantly over time. Originally, fallacies were defined as defective inferences or logically invalid types of arguments [40]. The pragma-dialectical theory of argumentation later redefined fallacies as *derailments of strategic maneuvering*, referring to speech acts that violate the rules of rational argumentative discussion for presumed persuasive gains [42, 41]. This perspective was further developed to describe fallacies as infringements of performance rules characteristic of a particular ideal type of argumentative engagement [43]. The concept was then refined to define fallacies as illicit dialectical

shifts across different dialogue types, emphasizing the inappropriate nature of argumentative moves in specific pragmatic contexts [165]. This evolution represents a significant shift in fallacy theory, moving away from a strictly rule-based, purely logical consideration of argumentative flaws towards a more nuanced, context-sensitive approach [169, 44]. As research in this field advanced, scholars began to recognize the complex, dynamic nature of real-world argumentation. This new understanding acknowledges that the validity and effectiveness of arguments must consider the specific context in which they occur, not just abstract logical principles. The modern approach to fallacies considers various pragmatic aspects, including participants' roles, dialogue purpose, shared knowledge, and broader social and cultural contexts [152]. This approach allows researchers to analyze why certain arguments may be fallacious in one context but potentially valid or persuasive in another. These insights are particularly significant in political discourse, where informal fallacies are strategically employed by politicians to advance their positions [95, 174, 172, 168]. Such deceptive strategic maneuvering can lead to faulty and biased reasoning by the audience and the formulation of further invalid arguments derived from those proposed by influential people, such as politicians.

2.4.1 Fallacies in Political debates

Political debates, with their high stakes and emotionally charged atmosphere, serve as a natural testbed for misleading arguments, offering numerous examples of fallacious reasoning that can persuade public opinion and influence critical decision-making processes. The study of these fallacious arguments is closely related to the field of propaganda detection, which has seen significant developments in recent years. A seminal work in this area is that of Da San Martino et al. [34], who developed an annotation scheme for propaganda methods that closely correspond to fallacious argument strategies. Their research resulted in a dataset of 7,485 annotated spans across 451 news articles, identifying 18 distinct propaganda methods. This dataset formed the basis for the NLP4IF'19 shared task on fine-grained propaganda detection, which analyzed content from 48 different news outlets. Building on this foundation, Da San Martino et al. [33] proposed a multi-granularity network architecture utilizing BERT contextualized embeddings. This approach allowed for the identification of propagandist content at various levels of granularity, including document, paragraph, and sentence levels. The significance of this work lies in its ability to provide a more nuanced understanding of how fallacious arguments and propaganda methods manifest in text. Further advancing the field, the SemEval 2020 Task 11 [32] refined the propaganda detection framework by reducing the number of categories and implementing a more stringent evaluation scheme. This task highlighted the overlap between propaganda categories and the definitions of fallacious arguments, underscoring the close relationship between these two areas of study. In the broader context of fallacy identification, various categorization schemes have been proposed. Some sources identify seven basic methods [116], while others expand this to at least 24 [171], and some discussions include up to

69 distinct types. For this analysis, we have focused on six main categories (as shown in Figure 2.3), drawing from the annotation scheme of [34] and the categorization proposed by Walton [167], which are particularly prevalent in political discourse. The importance of addressing fallacious arguments and propaganda methods cannot be minimized, given their potential to significantly impact societal decision-making processes. By understanding and identifying these methods, we can better equip both citizens and policymakers to critically evaluate the information they encounter, particularly in the context of political debates.

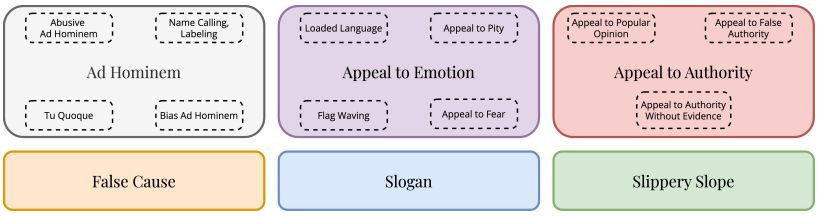


Figure 2.3: Fallacy categories for political debates.

2.4.2 Fallacy Classification

The pioneering work of Da San Martino et al. [34] laid the foundation for much of the subsequent research by introducing an annotation scheme of 18 propaganda methods applied to a dataset of 7,485 spans from 451 news articles. This dataset became a benchmark for fine-grained propaganda detection and served as the basis of the NLP4IF'19 shared task, stimulating further research in the field. Recent years have seen diverse approaches to fallacy classification. Jin et al. [81] proposed an architecture that incorporates the structural information of fallacies into a pre-trained language model, achieving an F1 score of 58.77% on a 13-class classification task. Building on this work, Alhindi et al. [5] introduced an innovative instruction-based prompt in a multitask configuration using the T5 model, identifying 28 distinct fallacies across various domains with F1 scores ranging from 17% to 62% across different datasets. Further advancing the field, Vijayaraghavan and Vosoughi [157] developed an end-to-end transformers-based approach for fine-grained propaganda classification in tweets, incorporating context and relational information to achieve a 64% F1 score on a 19-class classification task. This work demonstrates the potential of leveraging social media data for fallacy classification. It is important to acknowledge earlier works that paved the way for current research. Habernal et al. [55] developed "Argotario", an innovative gaming platform for fallacy annotation in everyday argumentation. Habernal, Pauli, and Gurevych [57] conducted experiments on fallacious argument type classification in German, achieving a 50.9% accuracy. Additionally, Habernal et al. [58] focused specifically on ad hominem fallacies, achieving an accuracy of 0.810 in predicting these arguments.

In my work, I advance the state of the art through a novel transformer-based model architecture fine-tuned on argumentation features. Through extensive evaluation, I demonstrate the necessity of detecting argument components and relations to improve

fallacy classification. My approach addresses the lack of resources for fallacious argumentation in political discourse, contributing not only a new resource but also an effective method for the largely unsolved task of classifying fallacious arguments in political debates. This work is motivated by the scarcity of existing resources in fallacious argumentation within political discourse and the need for effective methods to address this task.

In Section 6.1, my methodology advances fallacy classification through a novel transformer-based model fine-tuned on argumentation features. By incorporating argumentative features, this approach outperforms previous methods. This approach addresses the lack of resources for fallacious argumentation in political discourse, contributing not only a new resource but also an effective method for the largely unsolved task of classifying fallacious arguments in political debates.

The study of fallacies in political debates represents a culmination of advancements in natural language representation, Argument Mining, and fallacy analysis. By leveraging these interconnected fields, researchers can now develop sophisticated tools for identifying and addressing fallacious arguments in political discourse, contributing to more informed public debate and decision-making processes.

2.4.3 Fallacy Detection

Alongside classification efforts, researchers have developed comprehensive systems for fallacy detection. Vorakitphan, Cabrio, and Villata [161] proposed a two-step system that first detects potential propaganda and then classifies it based on the methods employed. This approach achieved a 72% F1 score for binary classification and a 64% micro F1 score for a 14-class classification task using standard benchmarks. Complementing this approach, Sahai, Balalau, and Horincar [139] focused on both comment-level and token-level fallacy classification, incorporating conversation context and using BERT-based models. Their work achieved a macro F1 score of 53% on token-level classification for 8 fallacy types, highlighting the importance of contextual information in fallacy detection. These detection systems demonstrate the progression from basic classification to more nuanced, multistep approaches that consider context and specific methods used in fallacious arguments.

The work in Section 6.2 advances fallacy studies mentioned above through three key innovations: (1) employing a corpus specifically for fallacy detection in political debates, (2) focusing on fallacy detection rather than classification, and (3) leveraging argumentation structure annotations. These contributions enhance the robustness and context-awareness of fallacy detection in political discourse analysis.

2.4.4 Fallacy Analysis with Large Language Models

Recent research has made significant progress in leveraging large language models to enhance various aspects of natural language processing, particularly in fallacy

recognition and factuality evaluation. Ruiz-Dolz and Lawrence [138] examined the use of LLMs for detecting argumentative fallacies, testing various models including SVM, fine-tuned RoBERTa, and zero-shot prompting with GPT-3.5 and GPT-4. Their results showed that while fine-tuned RoBERTa achieved the best performance (F1 score of 0.76), GPT-4 demonstrated competitive zero-shot capabilities (F1 score of 0.66). However, all models struggled with deeper logical understanding, indicating areas for future improvement. Alhindi, Muresan, and Nakov [6] demonstrated the power of LLMs in data augmentation for fallacy recognition tasks, using GPT-3.5 to generate synthetic examples addressing class imbalance issues. Their approach yielded consistent performance improvements across different fallacy types and datasets, showcasing the potential of LLMs in enhancing existing datasets and models. Glockner et al. [48] introduced MISSCI, a dataset for detecting fallacies in misinformation about scientific publications. Their evaluation of LLMs in a zero-shot setting showed promising results for GPT-4, which achieved a precision@1 of 0.317 for fallacy classification and generated plausible fallacious premises in 86.7% of cases according to human evaluation. However, LLaMA 2 struggled with the task, highlighting its difficulty and the need for further research in this area. Helwe et al. [69] presented MAFALDA, a unified benchmark for fallacy detection and classification. Their evaluation revealed that zero-shot detection was somewhat feasible (F1 scores ranging from 0.3 to 0.5 for different models), but fine-grained classification remained challenging for both LLMs and humans. This work underscores the complexity of fallacy analysis and the current limitations of even advanced language models. To address these challenges, Li et al. [99] proposed tasks from cognitive dimensions to enhance logical fallacy understanding. Their fine-tuned LLMs showed consistent improvements in logical reasoning tasks, with performance gains of up to 10% in accuracy compared to models trained on original data alone. Li et al. [100] introduced the FLUB dataset to test LLMs' understanding of fallacies using cunning questions. Their evaluation showed that current models struggle with fallacies and cunning questions, with even the best-performing models achieving accuracy rates below 50% on most tasks. Payandeh et al. [126] evaluated the reasoning capabilities and susceptibility to logical fallacies of GPT-3.5 and GPT-4 using the LOGICOM benchmark. Their findings revealed that LLMs are more easily convinced by fallacious arguments than logical reasoning, with GPT-4 being more susceptible than GPT-3.5. Specifically, GPT-4 agreed with 76% of fallacious arguments compared to 66% for GPT-3.5. Chiang and Lee [25] proposed D-FActScore, an enhanced metric for evaluating factuality in long-form text generations from LLMs. Their experiments with the AmbigBio dataset revealed that open-source LLMs like Llama and Tulu tend to mix information about distinct entities, resulting in D-FActScores 8-16% lower than their FActScores. In contrast, ChatGPT demonstrated a superior ability to disambiguate entities, with its D-FActScore only 1.5% lower than its FActScore.

Despite these advancements in fallacy detection and reasoning, existing works primarily focus on detection and classification rather than repairing fallacies, as shown in

Chapter 7. As far as current research indicates, no other work has employed these evaluation methods in the context of addressing fallacious arguments in political debates. By adapting these methods to the unique challenges of this domain, I aim to establish a benchmark for future research and contribute to the advancement of evaluation methods tailored to repair fallacies in political debates.

2.4.5 Related Models for Political Analysis

While not directly focused on fallacy detection, several models have been developed for related tasks in political analysis, which could potentially be adapted or integrated into fallacy detection systems: Liu et al. [104] introduced POLITICS, a pretrained language model for ideology prediction and stance detection. Leveraging novel pretraining objectives and a large-scale political news dataset, POLITICS outperforms baselines on 8 out of 11 tasks across 8 datasets. It achieves an overall average F1 score of 67.66, 3.6% better than RoBERTa, with notable improvements of over 10% on Hyperpartisan and YouTube user-level ideology labeling tasks. The model demonstrates effectiveness in handling longer documents, few-shot learning, and capturing salient political entities and sentiments. Hu et al. [73] introduced ConfliBERT, a pre-trained language model for political conflict and violence analysis. Evaluated on 12 datasets and 18 tasks, it consistently outperforms BERT, particularly with limited data. ConfliBERT maintains high F1 scores (65-73%) with as few as 34 training examples, while BERT's performance drops significantly. It excels in tasks like named entity recognition and multi-class classification, making it valuable for analyzing political conflicts and related fallacies. Jiang, Ren, and Ferrara [80] introduced Retweet-BERT, a model combining retweet networks and profile language to detect Twitter users' political leanings. Achieving 96%-97% macro-F1 scores on large datasets, it outperforms existing methods, especially with limited data. The model revealed significant right-leaning echo chambers in COVID-19 discussions. Retweet-BERT's performance suggests valuable applications for analyzing the ideological spread of fallacies and misinformation on social media.

2.5 Summary

This chapter traces the evolution of Natural Language Processing approaches, from fundamental language representation to advanced fallacy detection in political debates. It describes the core principles of Argument Mining, with a particular focus on fallacious reasoning in political rhetoric and its primary taxonomies. The discussion begins by exploring the path of NLP advancements that led to the development of AM. It then examines recent developments in AM, including classification methodologies, detection algorithms, and applications of large language models. Special attention is given to fallacy studies within the context of political discourse, highlighting the unique challenges and opportunities in this domain. Furthermore, the chapter contextualizes this

thesis's contributions within the current AM research landscape, emphasizing its significance and innovative aspects. By establishing the theoretical and methodological foundations, it sets the stage for the subsequent investigations and underscores the potential impact of this research on the fields of fallacy studies and political discourse analysis.

Chapter 3

Curation of ElecDeb60to20 and FallacyFix Datasets

This chapter introduces the ElecDeb60to20 dataset, an expanded collection of U.S. presidential debate transcripts from 1960 to 2020 annotated with argumentative components and structures [63] and FallacyFix that includes repaired fallacies from political debates, both created in the context of this thesis. ElecDeb60to20 features three annotation layers. The first layer identifies argument components, namely claims and premises. The second layer maps argument relations, focusing on attack and support relation between components. The third layer highlights fallacious arguments, which are speech acts that violate the rules of rational argument discussion for assumed persuasive gains. Initially covering debates from 1960 to 2016 [63], the final version includes the 2020 Biden vs. Trump debate. The second dataset, FallacyFix, comprises repaired versions of identified fallacies, to demonstrate examples of repaired fallacies commonly found in political debates. These comprehensive datasets serve as the foundation for all experiments in this thesis, offering insights into the evolution of political discourse. By analyzing argument structures and fallacies across six decades of presidential debates, ElecDeb60to20 and FallacyFix datasets provide a unique resources for understanding persuasion tactics in high-stakes political discussions.

The systematic extraction of argumentative structures from textual data requires a comprehensive corpus of annotated samples, which is crucial in the development, training, and empirical evaluation of machine learning classifiers employing diverse methodological approaches. To address this gap, Haddadan, Cabrio, and Villata [63] created the first version of a comprehensive annotated dataset of political debates, known as ElecDeb60to16. This dataset includes annotations for different argument components (claims and premises), argument relations (attack and support), and fallacious arguments. The initial version of ElecDeb60to16 comprised 54,640 components (29,004 claims and 25,635 premises), 25,012 relations (21,289 support and 3,723 attack), and 1,628 fallacies, derived from 41 debates, including 8 vice-presidential debates from 1960 to 2016. This dataset served as the foundation for the first line of

Transformer-based experiments, as detailed in Chapters 4, 6 and 5. As the initial version, ElecDeb60to16 immediately established itself as a unique resource, unique in its scope of annotated arguments and fallacies. To ensure the dataset's continued relevance and comprehensiveness, a second annotation phase was performed in the context of my thesis. This phase incorporated the final debates of the 2020 U.S. presidential election between Biden and Trump, adding 1,038 components, 512 relations, and 232 fallacious arguments to the corpus. This expansion was motivated by the need to keep the dataset current and reflective of contemporary political discourse. The ElecDeb60to20 dataset is composed of the following three layers of annotations:

- **Argument Components**: It consists of *claims* and *premises*. Claims represent the main point of an argument, often policy proposals in political debates. Premises are supporting statements that justify these claims.
- **Argument Relations**: The relations link components to create the argument's structure, resembling a graph. These relations can be supportive, or adversarial, reflecting how components interact with one another within the more general argument framework.
- **Fallacies**: Fallacies are rhetorical techniques that deviate from logical reasoning principles, employed with the intention of enhancing persuasive impact at the cost of argument integrity.

Additionally, I created **FallacyFix** that represents a subset of the dataset with fallacious annotations, in which selected examples have gone through fallacy repair methodology. It includes generated text that eliminates fallacious reasoning from the original fallacious arguments in political debates.

The subsequent sections of this chapter provide a comprehensive overview of the ElecDeb60to20 and FallacyFix datasets. Section 3.1 delineates the nature of the data, specifically U.S. Presidential Political Debates, and breaks down the collection methodology. Section 3.2 offers an in-depth examination of the various annotation phases and their respective protocols. The inter-annotator agreement (IAA) for all tasks, along with some analyses of disagreement, is presented in Section 3.3. The chapter concludes with Section 3.4, which provides detailed statistical insights into the ElecDeb60to20 and FallacyFix datasets, offering a quantitative perspective on its composition and scope.

3.1 Data Collection

This section presents a comprehensive exploration of the datasets' core content. The data collection process undergoes comprehensive analysis, clarifying each phase of corpus development. Furthermore, the section delineates precise specifications for both version of the ElecDeb60to20 dataset, as well as for the FallacyFix dataset.

The U.S. presidential election debates have been a source of both excitement and controversy since their very beginning. The two major political parties, the Democrats,

and the Republicans, compete vigorously during these debates to convince undecided voters, while supporters of each party eagerly anticipate watching their candidate engage in heated discussions over the most pressing issues of the day. Televised presidential debates have been a key component of American politics since 1960, when the first such debate between Kennedy and Nixon reached an astonishing audience of over 60 million viewers. This massive response underscores the crucial role these debates have played in shaping the country's political landscape. Over the years, the substance of these debates and the public's reactions to them have been extensively studied by political scientists, sociologists, and media analysts alike [85, 37]. The motivation for creating a new corpus is derived from two key factors:

1. To current knowledge, the dataset introduced by Haddadan, Cabrio, and Villata [63] remains, to our knowledge, the only existing corpus annotating political debates for both component and relational levels of argument structure. This unique status provided the rationale for expanding and enriching the dataset with recent debates and additional annotations, specifically to enable more complex analytical tasks such as fallacy detection.

2. Ensure the reproducibility of the annotation process by developing clear guidelines, inspired by [101, 136], that provide precise rules for identifying and segmenting argument components (i.e., claims and premises), argument relations (i.e., supports and attacks), fallacies, and methodologies to repair some of them, within the context of political debates.

The transcripts of televised debates between major presidential and vice presidential candidates were obtained from the official website of the Commission on Presidential Debates[1]. The initial release of the dataset, named ElecDeb60to16, spans from the 1960 debates between Kennedy and Nixon to the 2016 face-off between Clinton and Trump. I have updated this dataset to include the most recent debates from the 2020 election cycle, including both the presidential and vice presidential debates between Biden and Trump. The ElecDeb60to20 corpus includes 44 debates as detailed in Table 3.1. The primary characteristics that set this dataset apart are its vast scope, its focus on debates between the two dominant American political parties (Democrats and Republicans), and its organization along a temporal axis.

3.2 Annotation

The annotation process involves carefully analyzing the raw debate transcripts and identifying relevant argument relations, components, and fallacies within the text. This includes labeling specific statements as claims or premises, identifying the relations between these components (e.g., support or attack), and identifying specific categories of fallacies in political debates.

[1] https://www.debates.org/

Year	Types	Candidates
1960	4 pres	Kennedy - Nixon
1976	3 pres	Carter - Ford
1980	2 pres	Anderson - Carter - Reagan
1984	2 pres + 1 vice = 3	Mondale - Reagan
1988	2 pres + 1 vice = 3	Bush - Dukakis
1992	3 pres + 1 vice = 4	Bush - Clinton - Perot
1996	2 pres + 1 vice = 3	Clinton - Dole
2000	3 pres + 1 vice = 4	Bush - Gore
2004	3 pres + 1 vice = 4	Bush - Kerry
2008	3 pres + 1 vice = 4	McCain - Obama
2012	3 pres + 1 vice = 4	Obama - Romney
2016	3 pres	Clinton - Trump
2020	2 pres + 1 vice = 3	Biden - Trump
Total	**35 pres + 9 vice = 44**	

Table 3.1: Distribution of the presidential and vice presidential debates among the years.

The development of the annotation guidelines [61] was a collaborative process. A team consisting of three specialists in Argument Mining developed these guidelines for the annotation of argument components, relations, and fallacies, establishing a robust framework for the entire annotation process. In extending the corpus to include the 2020 debates, I adhered to these established guidelines [2].

3.2.1 Argument Components

The overall structure of an argument can be divided into two main parts. The first part consists of statements, which are the claims or assertions being made. The second part includes elements that support or validate these statements. These supporting elements are referred to as premises. This framework provides a way to isolate and analyze the components of an argument. The components are annotated using the following criteria:

- Annotators identify and mark the boundaries of claims and premises within the text. In this dataset, arguments typically lack major claims or explicit stances. The argument flow usually begins with the moderator's question, followed by candidates' alternating rebuttals. While major claims might be implicitly present in discussions of controversial issues like capital punishment or gun control, they are not explicitly annotated due to their infrequency.

[2] https://github.com/pierpaologoffredo/ElecDeb60to20/tree/main/guidelines

In this section, we detail the annotation of the argument components through some examples[3] from the USElecDeb60To20 dataset.

Claims

In argument discourse, claims represent the central objectives or conclusions. Within the realm of political debates, these claims often manifest as political proposals put forth by parties or candidates. Such proposals require substantiation to obtain acceptance from the audience. For instance, in Example 3.2.1, President Bush defends his administration's choices by asserting the effectiveness of his policies. In Example 3.2.2, Vice President Nixon advocates for the government's policies, aligning with his official role. Additionally, claims in political debates may extend beyond political proposals to include evaluations or critiques of opposing candidates or parties.

Example 3.2.1. BUSH: My administration started what's called the Proliferation Security Initiative. Over 60 nations involved with disrupting the trans-shipment of information and/or weapons of mass destruction materials. And [**we've been effective**]. [*We busted the A.Q. Khan network. This was a proliferator out of Pakistan that was selling secrets to places like North Korea and Libya*]. [*We convinced Libya to disarm*].[4]

Example 3.2.2. NIXON: Senator Kennedy's position and mine completely different on this. [**I favor the present depletion allowance**]. [*I favor it not because I want to make a lot of oil men rich*], but because [*I want to make America rich*]. Why do we have a depletion allowance? Because [*this is the stimulation, the incentive for companies to go out and explore for oil, to develop it*].[5]

Claims can also include taking a position on controversial topics or expressing opinions on specific issues. For example, "I've opposed the death penalty during all of my life" is considered a claim. Discourse markers like "in my opinion" or "I believe" often signal claims stating opinions or judgments.

Premises

Premises are supporting statements used by debaters to justify their claims. A common premise type is the appeal to experience, where seasoned candidates argue that their claims are more credible due to their extensive background. This strategy allows experienced politicians to leverage their track record to strengthen their arguments against less seasoned opponents, as shown in Example 3.2.3.

Example 3.2.3. CARTER: [*Well among my other experiences in the past, I've - I've been a nuclear engineer, and did graduate work in this field*]. [**I think I know the - the uh capabilities and limitations of atomic power**].[6]

[3]Claims are marked in **bold**, Premises are written in *Italic* and the component boundaries in squared brackets [].

[4]Bush-Kerry, September 30, 2004.

[5]Kennedy-Nixon, October 13, 1960.

[6]Carter-Ford, September 23, 1976.

Premises often include statistics and numerical data as evidence to support claims. Additionally, examples are frequently used as premises, often introduced by phrases like "for example" or "for instance". These strategies provide concrete support for the debaters' arguments.

3.2.2 Argument Relations

The identification of intricate argument structures within data necessitates the annotation of relations-directional connections between components. These relations link argument components to create argumentation graphs that represent the argument's structure. While traditional Argument Mining approaches typically aim to construct a tree structure with a single root node [147], the method used for this dataset takes a more data-driven stance.

Each debate begins with one candidate responding to a question from the moderator, a panelist, or the audience. The discourse then evolves as the opposing candidate offers their response and rebuttals to the first candidate's arguments.

Within this framework, two levels of arguments can be defined:

1. **Intra-speech arguments**: These occur within a single speech by a candidate. They consist of claims and premises connected by support and attack relations.
2. **Inter-speech arguments**: These span across different speeches, where components from one speech may support or attack elements from another.

Debates are characterized by their dynamic temporal structure, wherein arguments undergo progressive development across a series of speeches, all of which converge on the central issue posed by the moderator. Consequently, the arguments presented in each speech are interconnected with those made by opposing candidates in their respective turns. In the subsequent paragraphs, it will be explored the various labels used in our analysis. Furthermore, some examples will be provided to illustrate these labels within the context of political debates.

Support relation

Support relation links two components from a supporting argument component to a supported argument component. The argument component can either be a claim or a premise on both sides.

$$Arg1 \xrightarrow{\text{support}} Arg2, \text{ i.e., } Arg1 \text{ supports } Arg2 \tag{3.1}$$

The initial annotation schema allowed for multiple premises to independently support a single claim. This structure was illustrated in Example 3.2.4, where three distinct premises supported one claim. Initial argument relation annotations revealed complexities that necessitated guideline revisions. The updated guidelines require that each component should support only one other component, with a new rule implemented

for potential multi-support cases: annotators now link to the nearest supported component in the debate sequence.

Example 3.2.4. NIXON: But let's not put it there; let's put it in terms of the average family. What has happened to you? We find that [*your wages have gone up five times as much in the Eisenhower Administration as they did in the Truman Administration*]$_{\text{Premise1}}$. What about the prices you pay? We find that [*the prices you pay went up five times as much in the Truman Administration as they did in the Eisenhower Administration*]$_{\text{Premise2}}$. What's the net result of this? This means that [*the average family income went up fifteen per cent in the Eisenhower years as against two percent in the Truman years*]$_{\text{Premise3}}$. Now, [**this is not standing still**]$_{\text{Claim1}}$.[7]

In certain instances, as described in Example 3.2.5, multiple premises collectively support a single claim. The presence of linked arguments [147] introduced additional complexity to the annotation process. To mitigate this, the guidelines were revised: separate support relations are now established for each premise linking to the target claim, thereby decomposing complex argument structures into more simpler components.

Example 3.2.5. NIXON: We often hear gross national product discussed, and in that respect may I say that [*when we compare the growth in this Administration with that of the previous Administration that then there was a total growth of eleven percent over seven years*]$_{\text{Premise1}}$; [*in this Administration there has been a total growth of nineteen percent over seven years*]$_{\text{Premise2}}$. [**That shows that there's been more growth in this Administration than in its predecessor**]$_{\text{Claim1}}$.[8]

Support relations between premises are also frequently observed in argument structures. Example 3.2.6 illustrates this phenomenon, where one premise is fostered by another through the use of a specific example. In this instance, the example serves to reinforce the initial premise, collectively aiming to emphasize the destructive capacity of nuclear weapons.

Example 3.2.6. CARTER: [**This is a formidable force**]$_{\text{Claim1}}$. [*Some of these weapons have 10 megatons of explosion*]$_{\text{Premise1}}$. [*If you put 50 tons of TNT in each one of railroad cars, you would have a carload of TNT - a trainload of TNT stretching across this nation*]$_{\text{Premise2}}$. [*That's one major war explosion in a warhead*]$_{\text{Premise3}}$. [*We have thousands, equivalent of megaton, or million tons, of TNT warheads*]$_{\text{Premise4}}$. [**The control of these weapons is the single major responsibility of a President, and to cast out this commitment of all Presidents, because of some slight technicalities that can be corrected, is a very dangerous approach**]$_{\text{Claim2}}$.[9]

Attack relation

Attack relation holds when one argument component is in contradiction with another argument component. In an attack relation from argument A to argument B, A is trying

[7] Kennedy-Nixon, September 26, 1960.
[8] Nixon-Kennedy, September 26, 1960.
[9] Carter-Reagan, October 28, 1980.

to refute B.

$$Arg1 \xrightarrow{\text{attack}} Arg2, \text{ i.e., } Arg1 \text{ attacks } Arg2 \tag{3.2}$$

Claims can potentially refute other claims made by opposing candidates. However, it's important to note that the argument structure does not always align precisely with the sentences presented in the debate. In contrast to the approach taken with support relations, the annotation guidelines for attack relations were designed differently. Annotators were instructed to identify and mark all possible attack relations between components. This approach allows for a single component to potentially attack multiple other components. The decision to adopt this more comprehensive annotation strategy for attack relations was informed by their relative under-representation in the dataset. This method aims to capture a more complete picture of the argument dynamics, particularly focusing on the less frequent but crucial attack structures.

Attack relations can extend beyond individual speeches, occurring between components from different candidates' arguments. Example 3.2.7 illustrates this inter-speech dynamic. In this instance, Biden presents a justification for changes to the Medicare program, citing increased enrollment as evidence of success. The opposing candidate, Ryan, then interrupts, challenging the premise of this argument. The counterargument questions the reliability of the statistics used, pointing out that they originate from the first candidate's own actuaries, thus potentially compromising their objectivity. This scenario demonstrates how attack relations can function to undermine the credibility of an opponent's argument, not just by disputing claims directly, but by casting doubt on the validity of the evidence presented.

Example 3.2.7. BIDEN: [*More people signed up.*]$_{\text{Premise1}}$
RYAN: [*These are from your own actuaries.*]$_{\text{Premise2}}$[10]

3.2.3 Fallacies

For the first analysis of fallacious reasoning in political discourse, I have expanded upon the initial ElecDeb60To16 dataset [63]. This updated version now incorporates annotations of fallacious arguments found within U.S. presidential election debates spanning from 1960 to 2016. At this moment, this expanded dataset stands as the most comprehensive collection of political debates annotated not only for argument components (such as claim and premise) but also for argumentative relations (including support and attack). Before to start the annotation process, it has been conducted a preliminary analysis into the arguments presented by candidates in the ElecDeb60To16 dataset. This exploratory study aimed to identify which types of fallacies, among those outlined in the annotation scheme of [34] and the categorization proposed by [167], were most prevalent in political discourse. Based on this initial exploration, six types of fallacies occurred with notable frequency in political debates were chosen. Consequently, we decided to choose to concentrate my research on these specific categories:

[10] Biden-Ryan, 11 Oct, 2012.

Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, and Slogan. It's worth noting that the first three fallacy types are further subdivided into more specific subcategories to allow for a more nuanced analysis.

Furthermore, considering the most recent presidential election between Trump and Biden occurred in 2020, I further expanded the dataset with the transcripts of the debates of this election campaign to include updated annotations, incorporating argumentative components such as *Claims* and *Premises*, as well as the relations between these components, i.e., *Support* and *Attack*, and fallacies. As a result of this annotation update, the dataset is renamed as ElecDeb60to20, reflecting the coverage of debates spanning from 1960 to 2020.

Thus, the following paragraphs will describe the fallacies alongside examples and their respective justifications, describing why each example falls into a specific category of fallacy. The fallacy will be highlighted in slate blue. This approach aims to provide a clear and comprehensive understanding of how these techniques are employed to manipulate public perception.

Ad Hominem

When the argument becomes an excessive attack on an arguer's position [167]. It occurs when an argument is directed at a person rather than their position, actions, or arguments. Instead of addressing the merits of the situation, administration, or strategy, the focus is shifted to attacking the other candidate's character or personality. This type of fallacy involves insulting or discrediting the individual, rather than offering a substantive critique of their ideas or actions.

Among several subcategories of this fallacy [58], based on the context, four of them are described as follows:

1. **General/Abusive Ad Hominem.** It involves a direct attack on the character of the opponent rather than addressing the content of their argument. This tactic aims to discredit the individual personally, diverting attention from the actual issues being discussed.

   Example 3.2.8. Typical politician. All talk, no action. Sounds good, doesn't work. Never going to happen. Our country is suffering because people like Secretary Clinton have made such bad decisions in terms of our jobs and in terms of what's going on.[11]

   Justification: The candidate is trying to attack the person by mentioning stereotypical behavior of politicians (all action, no talk). Thus, his premise is annotated as **Ad Hominem**.

[11] Trump, Trump-Clinton debate, September 26, 2016.

2. **Name Calling, Labeling.** It involves tagging the object of the propaganda with terms that evoke fear, hatred, or disdain, or conversely, with terms that elicit admiration and praise. This tactic manipulates the audience's emotions to influence their perception of the subject.

   Example 3.2.9. Manchin says Democrats acted like babies at the SOTU (video) Personal Liberty Poll Exercise your right to vote. [32]

   Justification: The use of the expression like babies makes this example an **Ad Hominem** because it is trying to diminish their opposite by labeling them as babies.

3. **Bias/Circumstantial Ad Hominem.** It involves attacking an opponent by implying they have a personal interest or stand to benefit from their position in the argument. This tactic seeks to discredit the opponent's argument by suggesting their stance is driven by self-interest rather than objective reasoning.

   Example 3.2.10. I happen to support that in a way that will actually work to our benefit. But when I look at what you have proposed, you have what is called now the Trump loophole, because it would so advantage you and the business you do.[12]

   Justification: Mentioning that the opposing candidate will take a personal advantage of the tax law they are supporting is an example of **Circumstantial Ad Hominem**.

4. **Tu Quoque.** It occurs when an arguer evades criticism by accusing the opponent of similar behavior, rather than justifying their own actions. This tactic, meaning "you also", shifts focus from the argument to the opponent's conduct.

   Example 3.2.11. The book you mentioned that Vice President Gore wrote, he also called for taxing – big energy taxes in order to clean up the environment.And now that the energy prices are high, I guess he's not advocating those big energy taxes right now.[13]

   Justification: In this example, the candidate is pointing out that their opponent is hypocritical about their position on energy taxes and not addressing with a premise why he himself is opposing the energy tax. This is thus an example of Tu Quoque which is an **Ad Hominem** fallacy.

Appeal to Emotion

The use of emotion to support an argument can be fallacious if the emotional appeal is irrelevant to the logical validity of the argument. This fallacy occurs when emotional manipulation is employed to influence the audience's response instead of providing substantive evidence or reasoning. Politicians often exploit various categories of emotions, such as fear, anger, sympathy, and pride, to enhance their arguments, such as:

[12] Clinton, Trump-Clinton debate, September 26, 2016.
[13] Bush, Bush-Gore debate, October 11, 2000.

1. **Appeal to Pity.** It may be an evasion of relevant considerations needed to make a decision on the issue. For example, in a criminal trial if the defense attorney bases his whole argument on an appeal to pity, it could be reasonable to criticize his argument for its failure to look at the evidence for the defendant's guilt or innocence. [167]

   Example 3.2.12. So gun laws are important, no question about it, but so is loving children, and character education classes, and faith-based programs being a part of after-school programs. Some desperate child needs to have somebody put their arm around them and say, we love you.[14]

   Justification: Instead of providing relevant premises against conducting gun laws, the candidate tries to appeal to the emotion of the audience to feel pity for the children who commit shooting in schools. Thus, the premise is annotated as **Appeal to Emotion** fallacy.

2. **Flag waving.** It is a propaganda technique in which the debater tries to appeal to a group of people by using arguments which contain emotions concerning nation, race, gender, political preference or in general a group, idea, or country.

   Example 3.2.13. In 1933, Franklin Roosevelt said in his inaugural that this generation of Americans has a rendezvous with destiny. I think our generation of Americans has the same rendezvous. The question now is: Can freedom be maintained under the most severe tack — attack it has ever known? I think it can be. And I think in the final analysis it depends upon what we do here. I think it's time America started moving again.[15]

   Justification: By constantly talking of American will, the candidate is trying to appeal to the patriotic emotion to imply that a move (away from the previous administration) is needed in the U.S. So his claims are identified as **Appeal to Emotion** fallacy because it is not explicitly mentioning why this change is needed.

3. **Appeal to Fear.** It involves seeking to build support for an idea by instilling anxiety or panic in the population regarding an alternative. By exploiting fear, the argument attempts to manipulate the audience's emotional response to gain acceptance or compliance, rather than relying on logical reasoning or factual evidence.

   Example 3.2.14. Well, I think it's terrible. If you go with what Hillary is saying, in the ninth month, you can take the baby and rip the baby out of the womb of the mother just prior to the birth of the baby.[16]

   Justification: The candidate is trying to put fear of the law which the other candidate is proposing by painting an image of a baby ripped out of their womb as a

[14] Bush, Bush-Gore debate, September 26, 2000.
[15] Kennedy, Kennedy-Smith debate, September 26, 1960.
[16] Trump, Trump-Clinton debate, October 19, 2016.

premise to justify why this abortion law is not good. Thus, they are committing an **Appeal to Emotion**.

4. **Loaded Language.** It involves the use of specific words and phrases with strong emotional connotations, either positive or negative, to influence an audience. Politicians employ loaded language to evoke emotional reactions and sway public opinion, rather than presenting rational arguments or evidence.

   Example 3.2.15. Well, I actually agree with that. I agree with everything she said. I began this campaign because I was so tired of seeing such foolish things happen to our country.[17]

   Justification: The word "foolish" has a negative connotation and is a loaded word which will put the premise used in a **Loaded Language** and thus the argument is considered fallacious. It is not Ad Hominem since it is not directed to an opposite candidate.

Appeal to Authority

In this category of fallacious arguments, politicians use opinion of experts as their evidence. Three types of fallacies are defined based on this method of argument:

1. **Without Evidence.** It occurs when a claim is asserted to be true solely based on the endorsement of an authority figure, without providing any supporting evidence or reasoning to justify the claim.

   Example 3.2.16. (...) if we suffer defeat in Iraq, which General Petraeus predicts we will, if we adopted Senator Obama's set date for withdrawal, then that will have a calamitous effect in Afghanistan and American national security interests in the region (...).[18]

   Justification: The candidate is basing his argument on a premise that a defeat in Iraq is inevitable just because an authority said so without providing any other premise, thus he is committing a fallacy of **Appeal to Authority**.

2. **False Authority.** When a false authority's opinion is used as an evidence to support a claim which is not that authority's field of expertise.

   Example 3.2.17. (...) But in the case of missile defense, Senator Obama said it had to be, quote, "proven". That wasn't proven when Ronald Reagan said we would do SDI, which is missile defense. And it was major – a major factor in bringing about the end of the Cold War. We seem to come full circle again (...).[19]

   Justification: The candidate is trying to compare how Ronald Reagan handled the missile crisis in his time with the current time. Which is appealing to an authority on the relevant field but not related to the current situation. Thus, the candidate is committing the fallacy of **Appeal to Authority**.

[17] Trump, Trump-Clinton debate, October 9, 2016.
[18] McCain, Obama-McCain debate, September 26, 2008.
[19] McCain, Obama-McCain debate, September 26, 2008.

3. **Popular Opinion, Ad Populum**. This fallacy covers instances of attempted reinforcement of political claims by referring to the fact that something is very popular, or the will of the people.

   Example 3.2.18. He can make any excuse he wants, but the facts are that we're reducing the number of uninsured percentage of our population. And as the percentage of the population is increasing nationally, somehow the allegation that we don't care, and we're going to give money for this interest or that interest and not for children in the State of Texas is totally absurd. Let me just tell you who the jury is. The people of Texas. There's only been one governor ever elected to back-to-back four-year terms, and that was me.[20]

   Justification: The candidate is using his selection as governor to Appeal to Popular Opinion. He is saying: "since the majority of people chose me as governor for a four-year term, the popular opinion is that I am a good governor thus I am a good governor (who conducts correct insurance laws)". Thus, the whole premise is annotated as **Appeal to Authority**.

False Cause

It is the fallacy that provides the conclusion that some event happens as a result of an earlier event. In general, drive to the conclusion that some event is a result of a situation just because it happened at the same time or after. The misinterpretation of the correlation of two events for causation [167] is classified as an inductive fallacy. Politicians tend to apply this technique when they affiliate the cause of an improvement to their party, or the failure to their opponent's party.

Example 3.2.19. During the years between World War I and World War II, a great lesson was learned by our military leaders and the people of the United States. The lesson was that in the aftermath of World War I, we kind of turned our backs and left them to their own devices, and they brewed up a lot of trouble that quickly became World War II. And acting upon that lesson in the aftermath of our great victory in World War II, we laid down the Marshall Plan, President Truman did.[21]

Justification: The candidate is implying that the cause of World War II happening after World War I was that American Army left the area of War. In this example, the candidate is implying that the World War II began in the aftermath of World War I because the American's left the field. This is an example of **False Cause**.

Slippery Slope

It suggests that an unlikely, exaggerated outcome may follow an act. The intermediate premises are usually omitted, and a starting premise is usually used as the first step leading to an exaggerated claim.

[20] Bush, Bush-Gore debate, October 11, 2000.
[21] Gore, Bush-Gore debate, October 11, 2000.

Example 3.2.20. Now, what do the Chinese Communists want? They don't want just Quemoy and Matsu; they don't want just Formosa; they want the world.[22]

Justification: This example assumes that the initial desire for specific territories will unavoidably lead to a global conquest, without offering a reasoned explanation or evidence for this extreme conclusion. This creates a misleading and exaggerated view of the Chinese Communists' intentions and ignores other possible interpretations or outcomes.

Slogan

A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals. It can appear to invoke the excitement and discourage the counterpart.

Example 3.2.21. (...) if it doesn't work, then we have strengthened our ability to form alliances to impose the tough sanctions that Senator McCain just mentioned. And when we haven't done it, as in North Korea – let me just take one more example – in North Korea, we cut off talks. They're a member of the axis of evil.[23]

Justification: Using the familiar term axis of evil to talk about North Korea in his claim is a fallacy of using **Slogans**.

Given this complexity, the thesis focuses on six specific categories of fallacies that are particularly prevalent and impactful in political discourse [172]. These selected fallacies not only represent some of the most common logical flaws in political argumentation but also offer rich opportunities for analysis within their contextual frameworks [43]. By concentrating on these key categories, this thesis aim to provide a more nuanced and practical examination of fallacious reasoning as it manifests in real-world political debates. This focused approach allows for a deeper exploration of how these specific fallacies operate within the unique dynamics of political argumentation, taking into account the pragmatic aspects and contextual factors that traditional fallacy identification methods often overlook [152]. Furthermore, by limiting the focus to these six categories, the study can develop more targeted and effective strategies for identifying, analyzing, and potentially countering these fallacies in political discourse [76]. Ultimately, this selective focus serves as a stepping stone towards a more comprehensive understanding of fallacies in practice, providing future research that can expand upon these findings and potentially address a broader range of fallacy categories in various argument contexts.

The raw and processed data are available for access and review in the associated GitHub repository, which can be found at `https://github.com/pierpaologoffredo/ElecDeb60to20`.

[22] Nixon, Kennedy-Nixon debate, October 13, 1960.
[23] Obama, Obama-McCain debate, September 26, 2008.

3.2.4 Repaired Fallacies

The ElecDeb60to20-fallacy dataset [50] served as the foundation for addressing the challenge of repairing fallacious arguments in political debates. To facilitate a comprehensive evaluation of various large language models' capabilities in this domain, as elaborated in Chapter 7, I developed a novel dataset: FallacyFix[24], which comprises repaired versions of identified fallacies, to demonstrate examples of repaired fallacies commonly found in political debates. There are no (yet) objective guidelines on how an argument can transition from fallacious to non-fallacious, and there is no (yet) true definition of what it means to repair a fallacy. Given the current absence of an objective definition for fallacy repair, I have conceptualized it as follows: the *repair* of arguments containing fallacious arguments consists into *a more transparent, impartial version of the arguments devoid of potentially persuasive rhetorical techniques*. It is important to note that due to the inherent complexity and variability of language, this repair process may involve a degree of subjectivity. The primary objective is to ensure that the revised statements are free from biased language and manipulative rhetoric characteristic of fallacious arguments [40, 43]. For instance, the *repairing* approach for a Flag Waving fallacious argument could be:

Example 3.2.22. *"That's dangerous, and it's provocative. And the mixed message, the ambiguities of U.S. foreign policy, are –* ***I believe, and Bob Dole believes, is causing not only problems for this country throughout the world, but particularly here at home.*** *And the type of changes that were made overnight in California caused very severe dislocations."*[25]

The repaired version is rewritten as follows:

Example 3.2.23. *"That's dangerous and it's provocative. And the mixed message, the ambiguities of U.S. foreign policy, are –* ***This is causing problems everywhere.*** *And the type of changes that were made overnight in California caused very severe dislocations."*

Given that the ElecDeb60to20-fallacy dataset contains multiple categories of fallacy with different levels of complexity, we first conducted a pilot study to evaluate the feasibility of the task on a few of them, also in terms of human annotation effort and time efficiency. To start with, I considered *Appeal to Emotion* and *Appeal to Authority* as the main categories to focus on. To refine the approach, specific subcategories were further selected for my study:

- From Appeal to Emotion:
  - Loaded Language
  - Flag Waving
  - Appeal to Pity
  - Appeal to Fear

[24] It is publically accessible here.
[25] Jack Kemp, 09/10/1996.

- From Appeal to Authority:
  - Appeal to Popular Opinion
  - Without Evidence
  - False Authority

For the pilot annotation phase, we chose to focus on these subcategories to investigate fallacy repair for two primary reasons: *(1)* existing research suggests that emotional appeals and authoritative references are more readily identifiable and categorizable by annotators due to their distinctive features [60]. These features often include charged language, appeals to sentiment, or citations of purported experts, making them easier to spot. *(2)* Other categories, such as *Ad Hominem* and *False Cause*, require more nuanced contextual understanding, significantly increasing the complexity and effort of annotation [44, 147, 59]. For instance, identifying a *False Cause* fallacy often requires background knowledge of the subject to determine whether a causal relationship is legitimately established. This deeper analysis makes the annotation process more time-consuming and potentially less reliable for these categories. By focusing on more straightforward categories initially, I could refine the annotation protocols and inter-rater reliability metrics more efficiently before dealing with more challenging categories. Moreover, the selected subcategories allowed for a more scalable annotation process, crucial for generating a sufficiently large dataset for robust model training. Research supports this approach, indicating that emotional appeals and authoritative references are more readily identifiable by annotators due to their distinctive characteristics [58]. Motivated by the experimental pilot phase, expert annotators conducted the main annotation process using a refined methodology to ensure high-quality results on a larger scale. In contrast to complex annotation guidelines such as [61], the approach was deliberately straightforward, leveraging the context of fallacious arguments to employ basic operations. I primarily employed simple techniques such as removal, subtraction, and simplification, as illustrated in Figure 3.1.

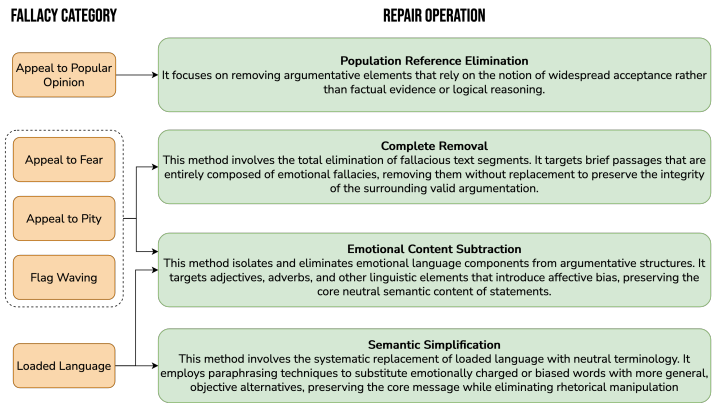


Figure 3.1: Diagram of methodologies developed and applied during the pilot annotation phase.

According to the proposed schema, *Appeal to Popular Opinion* often makes use of phrases implying widespread acceptance (e.g., "everyone knows"), that is targeted by **Population Reference Elimination**. Emotional fallacies like Appeals to Fear, Appeal to Pity, and Flag Waving commonly share traits of emotive language: *Appeal to Fear* appeals usually employ future-tense negative outcomes, *Appeal to Pity* appeals tend to use personal anecdotes and emotive adjectives, and *Flag Waving* generally relies on patriotic jargon. These may undergo **Complete Removal** if entirely emotional, or **Emotional Content Subtraction** if mixed with valid points. *Loaded Language*, typically marked by biased words and evaluative adjectives, can often be addressed through **Semantic Simplification**, replacing charged terms with neutral alternatives, or **Emotional Content Subtraction** to remove offensive, discriminatory, or inflammatory language while preserving core meaning. Each technique generally aims to preserve the argument structure while removing fallacious elements, guided by the specific syntactic features of each fallacy type.

To illustrate the methodology employed in repairing fallacies, Table 3.2 presents a selection of examples for the categories above based on the proposed annotation methodology.

Subcategory	Fallacious Argument	*Repaired* Argument
Appeal to Fear	The effort that we've mounted with respect to Iraq focused specifically on the possibility that this was the most likely nexus between the terrorists and weapons of mass destruction. **The biggest threat we faced today is the possibility of terrorists smuggling a nuclear weapon or a biological agent into one of our own cities and threatening the lives of hundreds of thousands of Americans.** What we did in Iraq was exactly the right thing to do.	The effort that we've mounted with respect to Iraq focused specifically on the possibility that this was the most likely nexus between the terrorists and weapons of mass destruction. *We faced hard situations*. What we did in Iraq was exactly the right thing to do.
Appeal to Pity	We just have a different approach. **But let me remind you, my family has suffered from drug abuse. I know what it's like to see somebody you love nearly lose their lives, and I hate drugs, Senator. We need to do this together and we can.** Not if I didn't have a better idea.	We just have a different approach. *We need to do this together and we can.* Not if I didn't have a better idea.
Appeal to Popular Opinion	No nation will ever have a veto over us. **But I think it makes sense, I think most Americans in their guts know, that we ought to pass a sort of truth standard.** That's how you gain legitimacy with your own countrypeople, and that's how you gain legitimacy in the world.	No nation will ever have a veto over us. *But I think it makes sense, that we ought to pass a sort of truth standard.* That's how you gain legitimacy with your own countrypeople, and that's how you gain legitimacy in the world.
Flag Waving	That will help reinforce the values that parents teach at home as well. **Ours is a great land, and one of the reasons why is because we're free.** And so I don't support censorship.	That will help reinforce the values that parents teach at home as well. *We live in a democracy.* And so I don't support censorship.
Loaded Language	And I was so surprised to see him sign on with the devil. But when you talk about apology, I think the one that you should really be apologizing for and the thing that you should be apologizing for are the 33,000 e - mails that you deleted, and that **you acid washed**, and then the two boxes of e - mails and other things last week that were taken from an office and are now missing. And I'll tell you what.	And I was so surprised to see him sign on with the devil. But when you talk about apology, I think the one that you should really be apologizing for and the thing that you should be apologizing for are the 33,000 e - mails that you deleted, and that *you handled in your own way*, and then the two boxes of e - mails and other things last week that were taken from an office and are now missing. And I'll tell you what.

Table 3.2: Examples of fallacious arguments repaired from the fallacy.

3.3 Inter-Annotator Agreement

Inter-annotator agreement (IAA) is a crucial measure in natural language processing and machine learning tasks that involve human annotation. It quantifies the degree of consensus among multiple annotators when they independently classify or label the same set of data. High inter-annotator agreement indicates consistency in the annotation process and suggests that the task guidelines are clear, and the categories are well-defined. On the other hand, low agreement may point to ambiguities in the task or inconsistencies among annotators, which can affect the reliability of the resulting dataset. To assess the level of consensus among annotators, I employed specific statistical measures for inter-annotator agreement (IAA). These measures were selected based on the annotation task's nature and the dataset's characteristics, following the methodology outlined by [146]:

- **Krippendorff's alpha**: This is a versatile reliability coefficient that can be used with any number of annotators, applicable to various types of data (nominal, ordinal, interval, ratio), and robust to missing data. It ranges from 0 to 1, where 1 indicates perfect agreement and 0 indicates agreement equivalent to chance. Krippendorff suggested that $\alpha \geq 0.667$ is the lowest conceivable limit for drawing tentative conclusions [86].
- **Observed agreement**: This is a simple measure that calculates the proportion of items on which annotators agree. It's calculated by dividing the number of agreements by the total number of items. While easy to compute and interpret, it doesn't account for agreement by chance [8].

3.3.1 Results of Argument Component Annotation

Three AM experts developed annotation guidelines, which were then applied by three different annotators, all non-native English speakers with high proficiency. Using the Brat annotation tool [148], at least two annotators independently annotated each transcript at the sentence level. The annotation was conducted at the sentence level, considering the presence and label of argument components in each sentence, without regard to exact boundaries. In the few cases (0.6%) where a sentence contained multiple components with different labels, the label of the longer component was used for analysis. To measure Inter-Annotator Agreement (IAA) for argument component detection, 22 out of 44 debates were independently annotated by three annotators. As reported in [63], for the task of distinguishing argument from non-argument sentences, an observed agreement of 83% and $\kappa = 0.57$ (moderate agreement) was achieved. For argument component classification, the observed agreement was 63% with $\kappa = 0.4$ (fair agreement). These results highlight the challenging nature of annotating political debates, particularly in differentiating between Premises and Claims.

3.3.2 Results of Argument Relation Annotation

The argument relation annotation phase followed the component annotation, involving three expert and three non-expert annotators in Argument Mining (AM). To manage the debates' complexity, they were divided into topic-based sections, typically marked by moderator questions. This structure guided the annotators in identifying argument relations within these contexts. Annotation quality was assessed through a two-stage Inter-Annotator Agreement (IAA) evaluation. The first stage, focusing on relation identification, showed high agreement (0.993 average, Fleiss $\kappa = 0.533$). The second stage, classifying relations as *Support* or *Attack*, proved more challenging (0.756 agreement, $\kappa = 0.387$), reflecting the nuances of political discourse.

The lower agreement on relation types led to a deeper analysis. It was discovered that one annotator diverged in quantity, not quality, of annotations. After careful consideration, their work was replaced with that of a more experienced annotator to ensure dataset reliability. This methodical approach has yielded a robust argument dataset, balancing comprehensive coverage with high reliability.

3.3.3 Results of Fallacy Annotation

The fallacy annotation process for political debates involved intensive training and continuous revision. For the first version[26], which included debates until 2016, annotators underwent a week-long preparation, studying guidelines and practicing on diverse debate segments. These discussions led to guideline updates, such as reclassifying **Appeal to Popular Opinion** under **Appeals to Authority** rather than **Appeal to Emotion**. Initial agreement analysis revealed issues with fallacy type boundaries, prompting further clarifications. To prevent bias, existing argument structure annotations were hidden from fallacy annotators. The annotation process utilized the INCEpTION platform [83] for data export, focusing on sentences across 10 documents, each annotated by three individuals with backgrounds in computational linguistics. The NLTK tokenizer package, with PunktParameters configured to handle common abbreviations like "dr, vs, mr, mrs, prof, inc", ensured accurate sentence segmentation [15]. To assess inter-annotator agreement, it has been employed two measures. The observed agreement [8] for sentences containing fallacies was notably high at 0.9655. However, to account for chance agreement, Krippendorff's α [86] has been calculated, which yielded a more moderate value of 0.4900. This dual approach provided a comprehensive assessment of inter-annotator reliability, balancing raw agreement with a chance-corrected measure to evaluate the consistency of fallacy identification across annotators. Following an initial annotation round on a sample dataset, it has been refined the guidelines to address disagreements, particularly regarding annotation span boundaries. To further evaluate inter-annotator agreement, it has been selected nine

[26] https://github.com/pierpaologoffredo/ElecDeb60to20/blob/main/data/fallacy_first_version.csv

sections from five debates spanning different years. The resulting agreement was moderate, as shown in Table 3.3. Subsequently, an expert annotator reconciled these annotations before incorporating them into the final dataset.

Fallacy Type	**Observed Agr.**	**Krippendorff's α**
Ad Hominem	0.9961	0.5315
Appeal to Authority	0.9945	0.5806
Appeal to Emotion	0.9759	0.4640
Slogans	0.9989	0.5995

Table 3.3: IAA, three annotators, 9 sections from 5 different debates (only 4 types of fallacies were present in the annotated data sample.)

For the second version of the dataset[27] that includes the 2020 debates of Biden vs. Trump, Two computational linguistics experts independently annotated argument components, relations, and fallacies. To ensure objectivity, argumentative components were annotated on raw data without prior fallacy annotations. Inter-Annotator Agreement (IAA) was evaluated using 50 randomly selected sentences of annotated fallacies, yielding substantial agreement: observed agreement of 0.857 and Krippendorff's α of 0.757. This process maintained annotation integrity while expanding the dataset to include recent political debates, enhancing its relevance for fallacy detection and classification in contemporary debates.

3.3.4 Results of Repaired Fallacy Annotation

Following up insights from the experimental pilot phase, two annotators with background in computational linguistics conducted the annotation process on a subset of the dataset following the methodology defined in the pilot study, to ensure high-quality results. The annotation of fallacious arguments was conducted on statements with the preceding and subsequent sentences relative to the speech turn containing the fallacy. This methodology leverages the context of fallacious arguments to produce repaired arguments. The repair process relies on simple editing operations such as removal, subtraction, and simplification, as shown in Figure 3.1. Both the original fallacious argument and the repaired one have been included into the **FallacyFix** dataset. This dataset was developed through a two-phase process. Two experts initially annotated a randomly selected set of 100 fallacious arguments from the ElecDeb60to20-fallacy dataset. To ensure objectivity and mitigate potential bias, the annotation process considered statements within their surrounding context. Specifically, annotators examined the preceding and subsequent sentences relative to the speech turn containing the suspected fallacy. This contextual approach was implemented to provide annotators with maximum relevant information for accurate fallacy repair. These arguments were drawn from the categories of fallacies under analysis. The experts followed the proposed methodology to annotate the arguments, allowing for the establishment of

[27] https://github.com/pierpaologoffredo/ElecDeb60to20/blob/main/data/fallacy_second_version.csv

inter-annotator agreement (IAA) between the two resulting repaired texts[28]. In this case, agreement was evaluated using two methods:

1. BERTScore [173]: Yielded a score of 0.94 ± 0.06
2. BERT [36] embeddings comparison: Resulted in a score of 0.98 ± 0.03

Additionally, to measure the extent of changes made during the repair process, we compared the human-repaired versions with the original fallacious statements using BERTScore. This comparison resulted in a score of 0.91 ± 0.06.. Given the current lack of standardized methods for comparing repaired fallacious arguments, BERTScore was selected as the most suitable metric for this task.

3.3.5 Disagreement

The next section delves into the disagreements noted among annotators and explores the challenges that emerged during the reconciliation process. This analysis sheds light on the areas where annotators diverged in their interpretations and the complexities faced when attempting to resolve these differences.

For the argument component annotation, to ensure annotation quality, it has been evaluated inter-annotator agreement (IAA) on a sample of 5 debates, comprising about 7,500 sentences. This evaluation compared the work of initial annotators against two experts in computational linguistics. Given the size of the dataset — 44 debates in total — it wasn't feasible to have multiple annotators for every debate. After refining the guidelines based on this evaluation, a single annotator completed the annotation of the remaining debates. For any disagreements in the sample set, it has been selected the annotations that aligned most closely with the experts' judgments for inclusion in the final dataset. A key factor contributing to annotator disagreement is illustrated in Example 3.3.1. The sentence beginning with "the way Senator [...]" serves as supporting evidence for a preceding claim. However, when read in a separate context, this same sentence could be interpreted as a claim itself.

Example 3.3.1. OBAMA: [**I disagree with Senator McCain in how to do it**]$_{\text{Claim1}}$, because [*the way Senator McCain has designed his plan, it could be a giveaway to banks if we're buying full price for mortgages that now are worth a lot less*]$_{\text{Premise1}}$.

Concerning the relation annotation, most of the disagreement was in assigning the relation label. In Example 3.3.2, two of the annotators have annotated $Claim_1$ as an attack to $Claim_2$, while the third annotator has annotated this relation as a support relation. According to the guidelines provided, the criticism on the assertions or policies suggested by the other candidate is considered as an attack relation. However, the choice of annotating this relation as a support can be explained from the assumption that this claim is the continuation of the chain of premises that support the claim about opposing tax increase.

[28] Each pair of compared texts was based on the same initial fallacious argument and fallacy category to be repaired.

Example 3.3.2. MCCAIN: [I will not stand for a tax increase on small business income]$_{\text{Claim1}}$. [*Fifty percent of small business income taxes are paid by small businesses*]$_{\text{Premise1}}$. [*That's 16 million jobs in America*]$_{\text{Premise2}}$. And **[what you want to do to Joe the plumber and millions more like him is have their taxes increased and not be able to realize the American dream of owning their own business]**$_{\text{Claim2}}$.[29]

Furthermore, disagreement was evident in relation classification, particularly when distinguishing between supportive and attacking claims. In Example 3.3.3, two annotators identified a relationship as supportive, while the third saw it as an attack. This disagreement arose from a claim's negative tone towards its subject, which seemed contradictory but actually reinforced the overall argument.

Example 3.3.3. FORD: It seems to me that **[in this election, the focus should not be on the executive branch]**$_{\text{Claim1}}$ but **[the corrections should come as the voters vote for their members of the House of Representatives or for their United States senator]**$_{\text{Claim2}}$. **[That's where the problem is]**$_{\text{Claim3}}$ and I hope there will be some corrective action taken, so we can get some new leadership in the Congress of the United States.[30]

The inter-annotator agreement (IAA) scores for both fallacy identification and repair tasks indicated a high level of consensus among annotators. In instances of disagreement, a structured resolution process was implemented. This process involved:

- Bilateral discussions between annotators
- Contextualization of individual perspectives, drawing from: *a.* Personal experience, *b.* Factual knowledge, and *c.* Interpretation of available information
- In-depth analysis of the text in question

This systematic approach to resolving inconsistency ensured a thorough examination of each contested annotation, ultimately leading to a refined and mutually agreed-upon classification. The high initial agreement and effective resolution process underscore the robustness of the annotation protocol and the reliability of the resulting dataset.

3.4 Dataset Statistics

To summarize, Table 3.4 reports on the total statistics of the argument components and relations annotation in the ElecDeb60to20 dataset, split by year. Concerning the annotation of the argument components, the analysis reveals a nearly equivalent number of premises and claims across all years, except for 1984. On average, there is a minimal difference of approximately -0.5% between the number of premises and claims. This close correspondence suggests a structured debate format where most claims are supported by specific premises. Regarding the annotation of relations, there is a significant

[29] Mccain-Obama, 15 Oct, 2008.
[30] Carter-Ford, 23 Sep, 1976.

imbalance between supporting and attacking relations. Supporting relations prevails over attacking relations by a factor of almost six to one. These findings provide insight into the debate structure:

- The near-parity between premises and statements indicates that candidates generally offer evidence or reasoning for their claims.
- The overwhelming preference for supporting relations over attacking ones suggests that candidates prioritize reinforcing their own arguments rather than critiquing their opponents' positions.

This pattern aligns with common political debate strategies, where candidates aim to present a strong, well-supported case for their own views while minimizing direct confrontations.

Year	Claims	Premises	Support	Attack
1960	1,858	1,724	1,231	205
1976	1,683	1,676	1,488	204
1980	449	440	653	115
1984	2,117	1,686	1,505	202
1988	2,291	1,571	1,583	323
1992	3,205	2,808	2,058	386
1996	2,752	2,634	1,926	333
2000	3,081	2,458	2,535	425
2004	3,526	3,362	2,506	566
2008	3,297	2,761	2,105	249
2012	3,778	3,712	2,424	428
2016	967	804	1,275	287
2020	620	419	400	112
Total	**29,624**	**26,055**	**21,689**	**3,835**

Table 3.4: Number of components and relations annotated in the final dataset, split by year.

Table 3.5 illustrates a clear distinction between intra-speech and inter-speech relations in the debates. Intra-speech relations, which occur within a single candidate's arguments, are predominantly supportive in nature. This is expected, as candidates typically use premises and claims to build and reinforce their own arguments. In contrast, inter-speech relations, which occur between different candidates' arguments, are markedly different. The majority (67%) of these relations are characterized as attacks. This stark difference exists because inter-speech relations often represent interactions between opposing candidates, where challenging and countering each other's positions is a common debate strategy.

Relation Type	Support	Attack
Intra-Speech	21,039	2,563
Inter-Speech	650	1,292
Total	**21,689**	**3,835**

Table 3.5: Final number and percentage of the annotated inter-speech and intra-speech relation types.

Table 3.6 presents the distribution of fallacy categories across 34 Presidential election debates[31]. Notably, the Appeal to Emotion fallacy consistently emerges as the most prevalent category throughout all years examined. The fallacious argument annotation phase covered 77.3% (34/44) of the initial corpus debates, constrained by resource limitations. Future research will extend annotation to the remaining 22.7%, enhancing dataset completeness and analytical reliability. The ElecDeb60To20-fallacy dataset offers valuable insights for political science scholars, particularly in analyzing the frequency of specific fallacious argument types. For instance, Table 3.6 highlights the prevalence of *Ad Hominem* arguments across various debate years. Of particular interest is the marked increase in the use of this strategy during the 2016 debates, suggesting a shift in debate tactics that year. Another noteworthy observation pertains to the 2004 debates, which exhibit a higher percentage of Appeal to Emotion fallacies compared to other years. This increased prevalence may be attributed to the debates' primary focus on the Iraq War, a topic that naturally lends itself to fear-based rhetorical strategies. These findings provide valuable context for understanding how rhetorical strategies in political debates evolve in response to current events and changing political climates.

Year	Debates	Ad Hominem	Appeal to Authority	Appeal to Emotion	False Cause	Slippery Slope	Slogans	Total
1960	4	10	24	95	12	12	1	154
1976	3	5	8	42	4	4	4	67
1980	2	5	12	77	2	3	5	104
1984	2	3	13	35	3	3	3	60
1988	1	4	19	31	2	3	4	63
1992	2	11	19	74	8	3	2	117
1996	2	10	24	93	6	2	10	145
2000	4	8	25	140	5	8	11	197
2004	4	32	38	135	13	10	4	232
2008	3	7	21	67	4	1	2	102
2012	1	0	2	16	1	1	2	22
2016	3	93	29	211	9	7	16	365
2020	3	62	17	147	0	4	2	232
Total	**34**	**250**	**251**	**1,163**	**69**	**61**	**66**	**1,860**

Table 3.6: Distribution of annotated fallacious argument spans among different debate years.

[31]The fallacious argument annotation process covered 34/44 of the initial corpus debates, constrained by resource limitations. Future research will extend annotation to the remaining part, enhancing dataset completeness and analytical reliability.

Furthermore, Table 3.7 summarizes the Trump-Biden debates' annotations by fallacy category and argumentative features. Analysis of tokenized fallacious arguments revealed Slogans as the shortest (5.0 tokens on average) and Slippery Slope as the longest (20.5 tokens).

Category	**Freq**	**AvgTok**	**Arg. Feature**	**Freq**
Ad Hominem	62	4.6	Claims	1,513
Appeal to Authority	17	18.6	Premise	332
Appeal to Emotion	147	6.81	Support Rel.	400
False Cause	0	0	Attack Rel.	112
Slippery Slope	4	20,5		
Slogans	2	5		
Total	232	9,25		2,357

Table 3.7: Distribution of annotated fallacies per category and argumentative features of Biden vs. Trump's debates.

As shown in Table 3.8, we present an overview of the FallacyFix dataset: a total of 747 fallacies, each paired with its corresponding repaired version.

It is possible to observe a higher proportion of instances from the Loaded Language category, followed by examples of Flag Waving. The remaining repaired categories are also represented, albeit with lower frequency. This distribution offers insights into the relative prevalence of different fallacy types within this dataset. This observed pattern may reflect the frequency of these fallacy types in real-world arguments in political debates, or could be indicative of their relative ease of repair.

Subcategory	**Distribution**	**Freqency**
Appeal to Fear	8.2 %	61
Appeal to Pity	1.1 %	83
Appeal to Popular Opinion	5.4 %	40
Flag Waving	19.7 %	147
Loaded Language	55.7 %	416
Total	100 %	**747**

Table 3.8: Statistics of the FallacyFix dataset.

For the dataset partitioning, I employed a stratified splitting strategy based on fallacy categories rather than debate chronology. This methodological choice was motivated by the uneven distribution of fallacy types across debates. By stratifying the split according to fallacy categories, I ensured a balanced representation of each fallacy type in the training, validation, and test sets. Following this stratified splitting strategy, I broke down the data as follows: I opted for 597 examples (80%) for training, 75 (10%) for validation, and 75 (10%) for testing.

Table 3.9 presents a comprehensive overview of the FallacyFix dataset's word count statistics. The dataset contains 747 examples, with an average word count of 58.72 and

a standard deviation of 30.39, indicating significant variability in example length. The word counts range from a minimum of 10 to a maximum of 249, covering both brief and elaborate fallacious arguments. The median of 54 words, slightly lower than the mean, suggests a slight right-skew in the distribution. With 25th and 75th percentiles at 36 and 76 words respectively, half of the examples fall within this 40-word range. This diverse composition ensures a rich dataset suitable for comprehensive fallacy analysis and repair technique development.

Statistic	**Stats**
Number of Examples	747
Average Word Count	58.72
Standard Deviation	30.39
Minimum Word Count	10
25th Percentile	36
Median (50th Percentile)	54
75th Percentile	76
Maximum Word Count	249

Table 3.9: Word count statistics for the FallacyFix dataset.

3.5 Summary

This chapter presents two crucial datasets: ElecDeb60to20 and FallacyFix. For the annotation of argumentative components and relations in political discourse, I developed comprehensive guidelines. These were first applied to create ElecDeb60to16, which was subsequently expanded into ElecDeb60to20. This dataset comprises 42,718 annotated sentences from U.S. presidential debate transcripts (1960-2020), featuring argument components, relations, and fallacies. The FallacyFix dataset offers repaired versions of fallacious arguments. The chapter outlines the annotation process, inter-annotator agreement, and dataset statistics. These resources, surpassing existing corpora in size and annotation depth, form the empirical foundation for all thesis experiments and provide insights into six decades of political discourse evolution.

Chapter 4

Argument Component Detection & Relation Prediction

This chapter introduces an Argument Mining pipeline for political debates, focusing on argument component detection and relation prediction. It explores various methods, from SVMs to fine-tuned Transformer models with RNNs for component detection, and approaches relation classification as a sequence classification problem using different transformer models. The research provides a comprehensive analysis of these approaches, including results and error analysis, highlighting challenges in capturing nuanced argumentative structures, particularly in classifying attack relationships and handling short, impactful sentences.

As discussed in Chapter 2.2, Argument Mining primarily consists of two standard tasks: argument component detection and relation prediction/classification. While some research has explored end-to-end modeling of these tasks [45], they are typically addressed separately due to their distinct challenges and complexities. This thesis proposes a comprehensive AM pipeline that integrates both tasks, aiming to leverage their interdependencies for improved performance. Figure 4.1 illustrates the proposed two-stage pipeline:

1. **Component Detection**: This initial stage focuses on detecting argument components within the input natural language text, specifically a political debate. Advanced neural networks, such as transformers or BiLSTMs with attention mechanisms, are employed to identify and delineate argument components (e.g., claims, evidence), with particular attention to component boundaries.

2. **Relation Classification**: The second stage predicts the relationships between the identified arguments, specifically attack and support relations. Additionally, in structured argumentation, this stage determines the internal relations between argument components, such as the connection between evidence and claims [147].

Experiment to test the proposed pipeline are run over the ElecDeb60to20 dataset, taking advantage from the argumentative annotations, discussed in Sections 3.2.1 and 3.2.2. The methodologies for detecting argumentative components and their relations

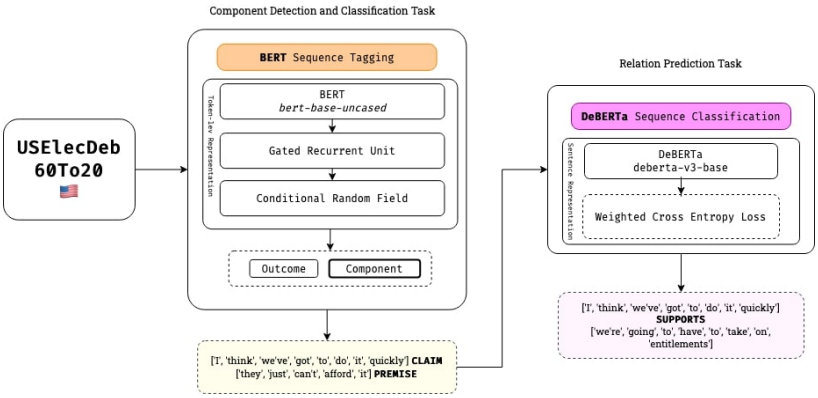


Figure 4.1: Illustration of the Argumentative Mining pipeline on political debates.

are thoroughly examined in Sections 4.1 and 4.2, respectively. These sections offer a critical evaluation of the results obtained from diverse approaches, emphasizing their efficacy and limitations in capturing the intricate relationships within argumentative discourse.

All source code and related resources for this research can be accessed on GitHub at `https://github.com/pierpaologoffredo/argumentation-mining-transformers`.

4.1 Argument Component Detection

The identification of argument components is commonly approached as a supervised learning task in text classification. This process involves a corpus of sentences, each annotated to indicate the presence or absence of an argumentative element. The objective is to develop a Machine Learning classifier capable of recognizing argumentative content within sentences. In formal terms, given a dataset $D = \{(x_i, y_i)\}_{j=1}^{N}$, where x_j represents a sentence and y_j its corresponding annotation (denoting whether the sentence contains an argument), the aim is to construct a discriminative function $f : X \to Y$ that can deduce the appropriate label from the input text. This challenge can be tackled using various Machine Learning techniques [101]. To incorporate the detection of component boundaries, the task can be reframed as a sequence labeling problem. In this formulation, x_j is not a single sentence with a unitary label, but rather a series of tokens, with y_j representing the sequence of corresponding labels. These labels correspond to the BIO (Beginning, Inside, Outside) tagging convention, indicating for each token whether it marks the start, continuation, or exclusion from an argumentative component. Sequence labeling tasks are often addressed using machine learning, and neural networks with recurrent or attention-based architectures. Section 4.1.1 and 4.1.2 provide an in-depth discussion of representative models for both the classification and sequence labeling approaches to argument component detection.

Word Embeddings Input word representations for sequence modeling can be created through two main methods: static lookup from pre-trained embeddings or dynamic generation using context-aware Language Models [130, 36]. Static embeddings offer simplicity but lack contextual understanding and may have limited vocabulary. In contrast, dynamic embeddings provide context-sensitive representations but require more computational resources.

Static Embeddings Static word embeddings are commonly used in natural language processing, with popular choices including GloVe [129], extvec [84], fastText [115], and BPEmb [68]. GloVe (100 dimensions) uses global word co-occurrence statistics, while extvec (300 dimensions) incorporates dependency graph information. Both are trained on large corpora like Wikipedia. To address out-of-vocabulary issues, especially in specialized domains, sub-word level embeddings are employed. FastText (300 dimensions) uses character n-grams and position weights, while BPEmb (100 dimensions) applies iterative merge operations on frequent symbols.

Dynamic Embeddings Dynamic word embeddings, such as ELMo [130], FlairPM [4], and BERT [36], generate context-aware word representations. ELMo uses bidirectional LSTMs to create contextual embeddings, while Flair employs character-based language models. BERT, based on a transformer architecture, jointly learns bidirectional representations considering sub-words and word positions. The BERTbase model, pretrained on BooksCorpus and Wikipedia, encodes words into 768-dimensional vectors.

Recurrent Neural Networks Sequence tagging assigns labels to each token in an input sequence, typically using Recurrent Neural Networks (RNNs) to model temporal dynamics. RNNs process sequences sequentially, using hidden states as memory. However, traditional RNNs suffer from short-term memory due to information loss over longer distances. To address this limitation, gated RNN architectures like Long Short-Term Memory (LSTM) [70] and Gated Recurrent Unit (GRU) [26] were developed. LSTMs use three gates (forget, input, and output) and maintain separate cell and hidden states. GRUs, a newer architecture, employ two gates (reset and update) and combine memory and hidden state functions. Both architectures aim to regulate information flow and preserve relevant long-term information. The BIO-tagging scheme is used to encode label information, where ideally a B-token (beginning) is followed by I-tokens (inside). This structured prediction task requires modeling token classifications dependently. Statistical graphical models, particularly Conditional Random Fields (CRFs) [88], are employed to represent the multivariate probability distribution and infer the most probable sequence of labels. In the context of this thesis, CRFs are used in conjunction with RNNs to enforce structured predictions. CRFs, viewed as a sequential extension of the Maximum Entropy model, consider predicted labels from other time steps to decode the most probable label sequence.

4.1.1 Baseline for Component Classification

In the context of component identification, this research explored sentence-level classification as an alternative to token-level classification. The sentence-level approach assigns a single label to entire sentences, indicating the presence of an argument component. For instance, a sentence containing a complete propositional statement would be labeled as a 'Claim'. This method contrasts with token-level classification, which requires predicting labels for individual words or tokens to precisely identify component spans and boundaries. Sentence-level classification provides a more streamlined approach compared to the comprehensive token-by-token examination traditionally necessary for identifying linguistic markers of argumentative components.

Experimental Setup The experimental methodology relied on two primary methods for the sentence-level classification approach. First, a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel was employed, utilizing Term Frequency - Inverse Document Frequency (tf-idf) as a feature. Regarding the SVM model, the number of epochs has been set to 3. Additionally, the research investigated the application of transformer models by fine-tuning a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model for sequence classification [36]. The model's hyperparameters were optimized using the Adam algorithm with a learning rate of 2e-5 over 3 epochs. A softmax activation function was applied to the output layer for binary classification. The pre-trained model and its associated tokenizer were instantiated using the Transformers library from HuggingFace[1]. These methods were selected to evaluate their efficacy in the context of argument component identification at the sentence level.

Results and Discussion The results obtained in performing the sentence-classification task are given in Table 4.1. Contrary to initial expectations, the Support Vector Machine (SVM) model, despite its relatively lower architectural complexity, demonstrated superior performance compared to BERT. The SVM achieved an accuracy of 49%, surpassing BERT's 44% by a statistically valuable margin of 5 percentage points. This unexpected outcome challenges the assumption that more sophisticated deep learning architectures invariably yield better results in this particular classification task.

Model	F1 Score
BERT	0.4400
SVM (tf-idf)	0.4900

Table 4.1: Results of the argument component detection task framed as sentence-level classification based on prediction of sentences among these labels {*Claim*, *Premise*, *Other*}.

[1] https://github.com/huggingface/transformers

Error Analysis An SVM model with tf-idf might outperform BERT in a multiclass classification task, such as labeling sentences as Claim, Premise, or Nothing in political debates, for several reasons. Firstly, tf-idf is effective when key words and their frequencies are the main discriminators, whereas BERT's complexity might be unnecessary for such tasks. SVMs also require less training data compared to BERT, which needs large datasets to capture nuanced linguistic contexts. Additionally, SVMs are less prone to overfitting with proper parameter tuning and feature selection, while BERT's complexity makes it more susceptible to overfitting without sufficient data and regularization. Furthermore, tf-idf directly captures relevant features that may be indicative of specific classes, such as keywords strongly associated with "Claim" or "Premise". The basic BERT, while contextual, might not prioritize these keywords correctly. Lastly, SVM with tf-idf involves simpler preprocessing and setup, whereas BERT requires careful fine-tuning to adapt to specific tasks. In summary, the reason why an SVM with tf-idf performed better is due to its simplicity, effective feature capture, lower data dependency, and reduced overfitting risk, making it well-suited for classifying sentences in political debates based on keyword importance.

4.1.2 Neural Network for Component and Boundary Detection

Consecutive experiments based on neural networks for the component detection task have been carried out. In the field of Argument Mining, many approaches focus on classifying component types while assuming predefined component boundaries. A novel approach integrates component classification and boundary detection into a single, unified task by recasting the problem as a sequence tagging task (illustrated in the first part of Figure 4.1). The approach employs a token-level schema based on the BIO (Beginning, Inside, Outside) tagging system. The possible label assignments for each token are drawn from the set: *B-Claim, I-Claim, B-Premise, I-Premise, O*. This fine-grained labeling allows for precise identification of argument components within the text. Traditionally, sequence tagging problems leverage the temporal dynamics inherent in text by utilizing a combination of Recurrent Neural Networks (RNNs) and Conditional Random Fields (CRFs). The proposed architecture builds upon this foundation while incorporating state-of-the-art natural language processing techniques. The core of the model is a pre-trained BERT model, fine-tuned for token-level sequence classification. This provides a robust baseline for understanding contextual relationships within the text. Following the BERT model, a Gated Recurrent Unit (GRU) layer [26] is implemented, which effectively captures sequential dependencies between tokens. The final component of the architecture is a Conditional Random Field (CRF) layer [88]. The interaction between these components is key to the model's effectiveness. While the GRU layer models the sequential nature of the text, the CRF layer exploits the inherent dependencies between labels in a sequence. This combination allows the model to jointly predict the most probable sequence of labels for all tokens, resulting in a coherent and context-aware labeling of the entire input text. By integrating these techniques, the approach aims to push the boundaries of argument component detection

and classification in AM tasks for the political debates.

Experimental Setup The sequence tagging experiments were conducted using three distinct architectural configurations: Transformer Model with Conditional Random Field (CRF), Transformer Model with Gated Recurrent Unit (GRU) and CRF, and Transformer Model in isolation. All Transformer models were implemented using the HuggingFace library (version 4.33.2) in PyTorch. To optimize model performance, a comprehensive hyperparameter search was conducted. The learning rate was selected from the set 1e-5, 2e-5, 3e-5, 4e-5, the number of epochs from 1, 2, 3, batch size from 8, 16, 32, and maximum sequence length from 32, 64. Based on this search, the optimal configuration for fine-tuning the BERT model was determined to be 3 fine-tuning epochs with an Adam optimizer, a learning rate of 3e-5, and a maximum sentence length of 32 tokens. This configuration was subsequently applied to the fine-tuning process for DeBERTa, DistilBERT, and XLM-RoBERTa models to ensure consistency across experiments. The dataset partitioning for training, evaluation, and testing remained consistent with the distribution outlined in Table 4.3, maintaining experimental continuity and allowing for direct performance comparisons across different model architectures.

Results and Discussion The results for the best-performing combination of models are shown in Table 4.2. Results are given in macro multi-class F1 Score and for claim and premise, respectively.

Model	F1	C-F1	P-F1
BERT + CRF	0.4712	0.4277	0.4201
BERT ForTokenClassification	0.4707	0.4279	0.4177
BERT + GRU & CRF	0.4684	0.4244	0.4167
DeBERTa + GRU & CRF	0.4674	0.4303	0.4098
DistilBERT + GRU & CRF	0.4605	0.4112	0.4088
XLM-RoBERTa + GRU & CRF	0.4553	0.4102	0.3972

Table 4.2: Results of the argument component detection task framed as token-level classification. Tokens are labeled with one of: {O, B-Claim, I-Claim, B-Premise, I-Premise}.

Generally, claim scores (C-F1) are higher than premise scores (P-F1) across all models, suggesting that claims are easier to detect than premises in political debates. One explanation for this phenomenon is that claims in political debates often follow certain rhetorical patterns or use specific persuasive language, making them more distinctive. Premises, on the other hand, can encompass a wide range of supporting information, facts, or examples, leading to greater diversity and complexity in their linguistic structure. This assumption is supported by the performance of the BERT ForTokenClassification model, which achieves results comparable to BERT + CRF. The effectiveness of BERT's built-in token classification capabilities suggests that there are significant claim-specific lexical and contextual patterns that the model can learn. Conversely, premises

in political debates can cover various topics and be presented in multiple ways, reducing the number of useful lexical cues for their detection. Another observation is that the performance of the models doesn't vary dramatically across different configurations, with F1 Scores ranging from 0.4553 to 0.4712. This consistency suggests that the base BERT architecture captures much of the necessary information for the argument component detection task in political debates. The addition of CRF layers provides a small but noticeable improvement, likely due to its ability to consider the entire sequence of labels rather than making independent predictions for each token. Interestingly, the incorporation of GRU layers doesn't significantly enhance performance, as seen in the BERT + GRU & CRF model's slightly lower scores compared to BERT + CRF. This aligns with the understanding that transformers, with their attention mechanism, are designed to capture long-distance dependencies without the need for recurrent architectures. In this case, it appears that BERT effectively captures the necessary contextual information for classification, making the additional sequence modeling of the GRU somewhat redundant. Comparing different BERT variants, we observe that specialized models like DeBERTa and DistilBERT don't lead to improved performance over the base BERT model in our political debate context. This contrasts with findings in other domains, where domain-specific models often outperform general ones. It suggests that the linguistic patterns in political arguments might not benefit as much from these specialized architectures as other domains do. The XLM-RoBERTa + GRU & CRF model shows the lowest scores across all metrics, indicating that multilingual models might not be optimal for this task, possibly due to the specific nuances and idioms present in political language that might not translate well across languages.

Error Analysis By analyzing the performance of the best model through the confusion matrix in Figure 4.2, we can make several key observations about the model's strengths and limitations in argument component detection. Misclassification of 'O' (non-argumentative) tokens is a prominent issue. The model frequently misclassifies 'O' tokens as either Claims or Premises. This error is particularly noticeable, with 53% of 'O' tokens misclassified as Claims or Premises (32% as B-Claim and 21% as I-Claim). This suggests that the model struggles to distinguish between argumentative and non-argumentative content, especially when similar vocabulary or semantic structures are used. Context plays a crucial role in argument identification, which the model may not fully capture. Political debates often contain implicit information that humans can infer but models struggle with, highlighting a limitation in the model's understanding of nuanced discourse.

There's a notable bidirectional confusion between Premises and Claims. 25% of B-Premise tokens are misclassified as "O", and 13% as B-Claim. Similarly, 19% of I-Premise tokens are misclassified as "O", and 13% as I-Claim. Conversely, 7.6% of B-Claim tokens are misclassified as B-Premise, and 1.1% of I-Claim tokens as B-Premise. This confusion indicates that the model struggles to differentiate between the argumentative roles of Claims and Premises. It also points to the inherent ambiguity in political

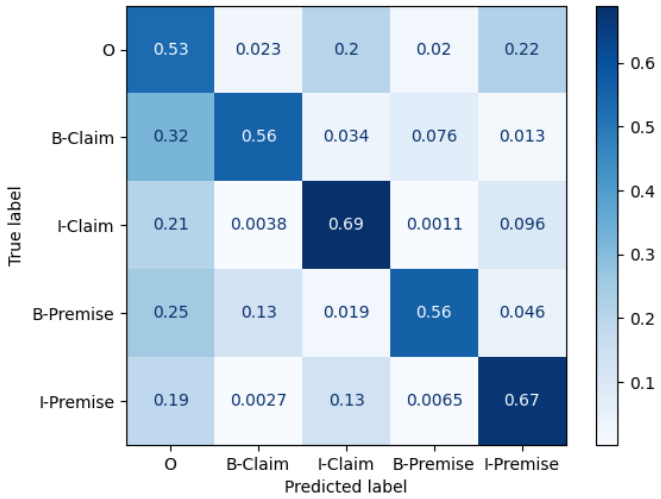


Figure 4.2: Confusion matrix of Component Detection and Classification of the best model (BERT + CRF) among all the labels.

discourse, where Claims and Premises can be closely intertwined or similarly phrased. A potential imbalance in the training data (as noted in Table 4.3) may be influencing the model's ability to distinguish between these components. The model shows difficulties in accurately identifying the boundaries of argumentative components. There's confusion between 'B-' (Beginning) and 'I-' (Inside) tokens for both Claims and Premises.

The seqeval [122] analysis revealing a macro F1 Score of only 32% for exact boundary matching (33% for Claims, 31% for Premises) further emphasizes this limitation. This suggests that the model struggles with the sequential nature of argument component identification and may not fully capture the structural patterns that differentiate the beginning of an argument component from its continuation. Despite these challenges, the model shows some strengths. It correctly identifies 69% of I-Claim tokens and 67% of I-Premise tokens, suggesting better performance in continuing to classify a component once it has been identified. The model performs reasonably well in identifying B-Claim (56% accuracy) and B-Premise (56% accuracy) tokens when it does recognize them as argumentative.

A few examples[2] of wrong predictions that exhibits these errors are listed here:

Example 4.1.1. Now, just because we don't want to get involved everywhere doesn't mean we should back off anywhere it comes up.$_{Other}^{Claim}$

Example 4.1.2. But if our national security is at stake, if we have allies, if we've tried every other course, if we're sure military action.$_{Other}^{Premise}$

Example 4.1.3. [*Too many people have been left behind.*]$_{Premise}^{Other}$

[2]We adopted the same representation of the components used in the Section 3.2.

Example 4.1.4. **[We've caught'em all.]**$_{Claim}^{Other}$

Example 4.1.5. [*The key is job training, education, investments in health care and education, environment, retirement security.*]$_{Premise}^{Claim}$

4.2 Argument Relation Prediction

After the argument component detection, the focus now shifts to the crucial task of predicting relations among political claims or premises (see second part of Figure 4.1). This next step in the argument mining process relies on the identified components to establish a more comprehensive understanding of the argumentative structure within political debates, as already mentioned in Section 2.2. After extracting valid sequences of Beginning-Inside (BI) tags from the component detection phase, a set of argumentative components for each political discourse is obtained. These components, which may be phrases rather than complete sentences, form the framework for the relation classification task. Relation classification in this context can be approached from various points of view. In the current research, it is framed as a sequence classification problem. This approach involves analyzing pairs of argumentative components to determine the relationship between them. Unlike methods that aim to construct predefined argumentation schemes, this approach classifies relationships between components independently, without imposing constraints of a predefined argument structure. To tackle this challenge, Transformer models are employed. These architectures have demonstrated great performance in tasks involving the classification of relationships between textual segments, making them particularly well-suited for the purpose of analyzing political discourse. By thinking of relation classification as a sequence classification task, two potential strategies are shown:

- A holistic approach that jointly models relations by classifying all possible combinations of argumentative components.
- A two-step method that first predicts potential link candidates for each component, followed by relation classification only for the most plausible pairs.

Both of these approaches have precedents in existing literature, each offering unique advantages in the context of political argument analysis. This methodology allows for the capture of nuanced connections between claims and premises in political debates, providing a more general picture of the argumentative context. By understanding these relationships, deeper insights into the structure and flow of political arguments can be gained, potentially revealing patterns in reasoning, persuasion techniques, and the overall coherence of political discourse.

In this thesis, a comprehensive evaluation of relation classification in political argumentation is conducted, beginning with fundamental approaches and moving on to more advanced methodologies. The research initiates with experiments where pairs of argumentative components are combined into single sentences, to which relationship

labels are assigned, performing the sequence classification task in its simplest form. Based on this starting point, the study then extends to a series of more complex experiments that investigate the efficacy of various Transformer architectures in addressing this multiclass problem. These advanced experiments are designed to assess how different Transformer models perform in capturing and classifying the relations between argumentative components in political discourse, thus providing a holistic approach to the challenge of relation classification in political argumentation.

Experimental Setup The model architecture employed for this task is based on bidirectional transformers [156]. This architecture comprises an encoder and decoder, each containing multi-head self-attention layers followed by fully-connected dense layers. Unlike the sequence tagging transformer, which processes individual token representations, the sequence classification task requires a pooled representation of the entire sequence. For relation classification, the input consists of a pair of components separated by a special token, rather than a single sentence as in sequence tagging. The pooled representation of this component pair is then fed into a linear layer with a softmax function, which generates a probability distribution over the target classes. This approach results in a three-class classification problem: Support, Attack, and NoRelation. A key advantage of this architecture is its flexibility; it allows for the possibility of one component having relations with multiple other components, as each component combination is classified independently. This feature is particularly valuable in the context of political discourse, where arguments often have complex, interconnected relationships.

DeBERTa [66] showed superior performance in generating token-level representations for contextualized sentences. The model was fine-tuned over 3 epochs using an Adam optimizer, with a learning rate of 4e-5 and a maximum sentence length of 256 tokens. This optimal configuration was consistently applied in the fine-tuning process for other models, including RoBERTa, DistilBERT, and XLM-RoBERTa. The hyperparameters set for BERT in the simple configuration are 3 epochs, learning rate of 2e-5, and maximum sentence length. Regarding the SVM model, the number of epochs has been set to 3. The dataset was strategically partitioned to support robust model training and evaluation. The training set comprised 40,838 relations, while 5,105 relations were allocated each to the evaluation and test sets.

In this case, this method involves creating comprehensive lists of all identified components for each dataset (training, evaluation, and testing). To establish the dataset for classification, each component is systematically paired with every other component within its respective set. These pairings are then labeled according to their pre-existing relationships: attack (r_a) or support (r_s). To ensure a balanced dataset, non-relationships (r_{nr}) are introduced such that their quantity equals the sum of attack and support relationships ($r_{nr} = r_a + r_s$), as illustrated in Table 4.3. The weight factor of the three classes in the (weighted) Cross Entropy Loss is taken into consideration, with normalization applied to the number of training samples in each class.

The debates were distributed across training, validation, and test sets with consideration for election years. In years featuring four debates, one was allocated to the validation set. This methodology ensured that each election year was represented by exactly one debate in the test set, maintaining a balanced distribution across temporal boundaries. The dataset statistics are presented in two comprehensive tables: Table 4.3 illustrates the distribution of component and relation types across the different sets.

Set	# Debates	Claims	Premises	Total %	Support	Attack	No relations	Total %
Train	23	22,883	20,656	78.2%	17,351	3068	20,419	80%
Validation	8	4,825	3,693	15.3%	2,169	383	2,553	10%
Test	13	1,916	1,706	6.5%	2,168	384	2,552	10%
Total	**44**	**29,624**	**26,055**	**100%**	**21,689**	**3,835**	**25,524**	**100%**

Table 4.3: Distribution of components and relation split into training, validation, and test sets.

Results and Discussion The results for relation classification are shown in Table 4.4. Results are given in macro multi class F1 Score. The study compared several state-of-the-art transformer-based models using a holistic approach, as well as two baseline methods employing a sentence classification (sent-class) approach. Among the models tested, DeBERTa emerged as the best model, achieving a macro F1 Score of 0.7032. This was followed closely by RoBERTa (0.6753) and BERT (0.6646), both utilizing the holistic method. The strong performance of these transformer-based models underscores their effectiveness in capturing complex relationships between argumentative components in political debates. Notably, XLM-RoBERTa and DistilBERT, while still employing the holistic approach, showed comparatively lower performance with macro F1 Scores of 0.6374 and 0.5817 respectively. This variance in performance among transformer models suggests that architectural differences and pre-training strategies play a significant role in their ability to classify relations in political argumentation. The comparison between holistic and sentence classification approaches yields interesting insights. The BERT model, when used in the holistic method, outperformed its sentence classification counterpart by a considerable margin (0.6646 vs. 0.5900). This disparity highlights the potential benefits of considering the broader context and structure of arguments in relation classification tasks. The traditional SVM with tf-idf features, representing a simpler machine learning approach, achieved the lowest macro F1 Score of 0.3400, further emphasizing the superiority of deep learning models in this complex task. These findings suggest that while simpler approaches like sentence-level classification can provide a baseline, more sophisticated methods that consider the holistic nature of argumentative structures are crucial for achieving higher performance in relation prediction within political debates.

Model	Method	Macro F1
DeBERTa	*holistic*	**0.6877**
RoBERTa	*holistic*	0.6753
BERT	*holistic*	0.6646
DistilBERT	*holistic*	0.5817
XLM-RoBERTa	*holistic*	0.6374
BERT	sent-class	0.5900
SVM (tf-idf)	sent-class	0.3400

Table 4.4: Results of Relation Prediction task based on sequence classification among the labels {*Support, Attack, NoRel*}.

Error Analysis The performance of the best relation prediction model (DeBERTa) reveals several notable patterns and challenges, particularly when considering the inherent class imbalance in the dataset and the nuanced nature of argumentative relationships in political debates. As illustrated in Table 4.3, there is a significant imbalance in the distribution of relationship types across the training, validation, and test sets. The 'No relations' category is the most prevalent, accounting for approximately 50% of the total instances, followed by 'Support' relationships at about 42.5%, and 'Attack' relationships comprising the remaining 7.5%. This imbalance presents a challenge for the model, potentially affecting its ability to accurately identify the less represented 'Attack' relationships.

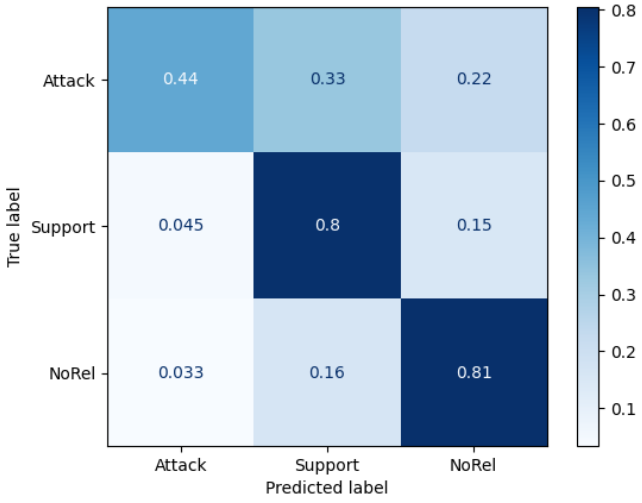


Figure 4.3: Confusion matrix of Relation Prediction task of the best model (DeBERTa).

The confusion matrix in Figure 4.3 provides valuable insights into the model's performance across different relationship types. The model demonstrates high accuracy in identifying "No Relation" instances, with a true positive rate of 0.81. This strong performance aligns with the class's prevalence in the dataset, but may also indicate a bias

towards this majority class. For "Support" relationships, the model shows good performance, with a true positive rate of 0.80. However, there is growing confusion with "No Relation" (0.15 false negative rate) and "Attack" (0.05 false negative rate) categories. The model struggles most with "Attack" relationships, achieving only a 0.44 true positive rate. Critically, 33% of actual "Attack" relationships are misclassified as "Support", and 22% as "No Relation". This poor performance on "Attack" relationships is likely due to the severe class imbalance and the implicit distinctions between "Attack" and "Support" in some contexts.

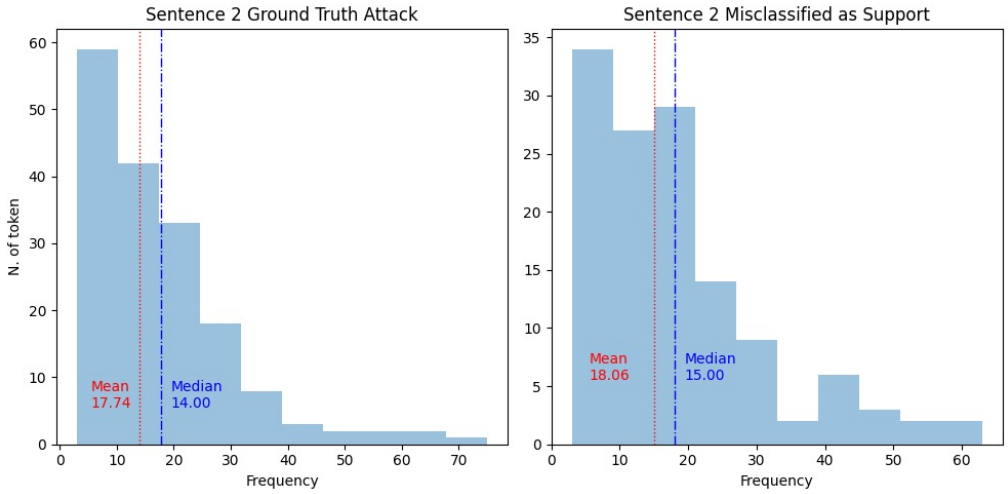


Figure 4.4: Analysis of the second sentence between true and misclassification of the *Attack* relationship.

Figure 4.4 provides an intriguing analysis of sentence length in relation to misclassification. The second sentences in correctly identified "Attack" relationships have a mean length of 17.74 tokens and a median of 14 tokens. Interestingly, the second sentences in "Attack" relationships misclassified as "Support" have very similar length characteristics, with a mean of 18.06 tokens and a median of 15 tokens. This similarity in sentence length distribution suggests that purely quantitative features like token count are insufficient for distinguishing between "Attack" and "Support" relationships. The model may be overly relying on surface-level features, failing to capture the deeper semantic distinctions between these relationship types. Beyond statistical measures, the analysis reveals important semantic and structural patterns in misclassified instances. Misclassified sentences often contain repeated phrases or expressions typically used to emphasize points, engage the audience, or reinforce idea connections. These rhetorical devices may confuse the model, particularly in distinguishing between supportive elaboration and subtle attacks. Additionally, short, semantically dense sentences carrying significant emotional weight are frequently misclassified. These often connect two premises rather than a premise and a claim, suggesting that the model struggles with complex argumentative structures.

A few examples[3] of wrong predictions that exhibits these errors are listed here:

Example 4.2.1. **[That's not the way this country ought to be run]** $\overset{\text{Attack}}{\underset{\text{Support}}{\Longrightarrow}}$ **[Mr. Ford, so far as I know, except for avoiding another Watergate, has not accomplished one single major program for this country]**

Example 4.2.2. **[If we remain dependent on a source of energy that is outside our control, we're not going to be as strong as we should be]** $\overset{\text{Attack}}{\underset{\text{Support}}{\Longrightarrow}}$ **[No matter how strong we are economically]**

Example 4.2.3. **[He has raised taxes several times]** $\overset{\text{Attack}}{\underset{\text{Support}}{\Longrightarrow}}$ [*The governor has to balance the budget in his state he is required to by law*]

Example 4.2.4. **[That's not right]** $\overset{\text{Attack}}{\underset{\text{NoRelation}}{\Longrightarrow}}$ **[The tax code is unfair for people at the bottom end of the economic ladder]**

The cases highlight the model's difficulty in handling brevity, context dependency, subtle contradictions, and the mixing of factual statements with criticisms. Short, impactful sentences, as seen in Examples 4.2.1, and 4.2.3. The model frequently struggles to connect brief statements with their more detailed follow-ups, missing the overall attacking nature of the relationship, as demonstrated in Examples 4.2.2 and 4.2.4. The model demonstrates significant challenges in recognizing implied criticisms and implicit attacks. It often fails to identify the attacking intent when criticisms are not explicitly stated and struggles to differentiate between objective statements and their strategic use as subtle attacks. This difficulty in capturing nuanced language and context leads to misclassifications, particularly when dealing with statements that appear factual on the surface but carry underlying critical intent, as shown in Example 4.2.3.

4.3 Summary

This chapter introduces a significant methodological advancement in the field of Argument Mining applied to political debates. The major contribution is a novel supervised approach for argument component identification and relation prediction within political discourse. This method employs a BERT-RNN-CRF architecture for component detection, achieving an F1 Score of 0.47, and a sequence classification model for relation prediction, attaining an F1 Score of 0.69. Notably, this approach surpasses standard AM baselines in performance. Further, a comprehensive linguistic analysis was conducted, revealing the unique challenges inherent in applying AM to political debates, such as the frequent occurrence of implicit statements and propagandistic assertions.

[3] It has been adopted the same representation of the components used in the Section 3.2.

Chapter 5

DispuTool

This chapter introduces DispuTool, an innovative system designed for the automated analysis of political debates from an argumentative perspective. Developed to identify argument components and their relations, DispuTool offers a sophisticated approach to visualizing debates as graphs. The system's capabilities extend beyond mere argument mapping, providing comprehensive data exploration features through entity recognition, fallacies, and analytical graphical representations. To demonstrate its practical applications, this chapter examines DispuTool's implementation in real-life political scenarios, specifically focusing on U.S. Presidential Debates. The major contribution underlying this chapter focuses on the updated version (2.0) of DispuTool, emphasizing its contribution to the field of computational argumentation and political discourse analysis.

This section presents the new version of DispuTool[1], a system designed to address the challenges involved in processing the large amount of textual data generated by political debates, with a particular focus on U.S. presidential elections. Political speeches and debates exemplify scenarios where large quantities of textual data require analysis to formulate or support hypotheses in historical and social scientific research. Political debates, specifically, function as public forums where electoral candidates engage in direct discourse on critical issues such as unemployment, taxation, and foreign policy. To validate the efficacy and utility of the proposed Argument Mining pipeline for political debates (see Chapter 4), a demo system, DispuTool, was implemented. The key features of DispuTool include the automatic identification and classification of argumentative components within debate transcripts and the prediction of argument relations between these components. Another feature of DispuTool is the integration of Named Entity Recognition (NER) in conjunction with argument mining and topic modeling techniques. This holistic approach allows for a contextualized exploration of argumentative structures and topics, facilitating the identification of speakers, referenced entities, and their interactions within the debate's argumentative framework. The technical architecture of DispuTool incorporates advanced machine learning algorithms and Neural Networks to perform the AM pipeline in political debates. These computational

[1] https://3ia-demos.inria.fr/disputool/

methods are crucial in executing the system's two primary tasks, as mentioned in Section 2.2: *(1)* Argument Component Detection and *(2)* Relation Prediction. Through the application of these methods, DispuTool effectively transforms complex debate data into structured, analyzable information. Thus, DispuTool represents a significant advancement in the analysis of political discourse, combining artificial intelligence with traditional humanities research methods. This innovative tool empowers historians, political scientists, and other researchers to conduct in-depth examinations of debate mechanics and persuasive strategies. By leveraging computational power, DispuTool reveals patterns and structures in political arguments that might go unnoticed using conventional analytical techniques. To our knowledge, DISPUTool is unique in its comprehensive approach. It not only identifies argumentative components and their relationships but also incorporates Named Entity Recognition (NER) and graph overviews. This combination enables intelligent exploration of political debate transcripts, offering a more holistic analysis than previous tools in this field. DispuTool contributes to the growing body of literature on using technology to study political debates, offering a fresh approach to argument mining. Its development illustrates how interdisciplinary collaboration between computer science and the humanities can yield powerful new research tools.

Section 5.1 introduces the initial version of DISPUTool 1.0Following this, Section 5.2 presents the updated (2.0) version [49], featuring additional capabilities with further enhancements in development.

5.1 DispuTool 1.0

The first version of DispuTool [62] was developed by a former PhD student within our team. This tool was designed to analyze American presidential debates, leveraging the ElecDeb60to16 dataset, which included debates that ranged from 1960 to 2016. Its primary function is to automatically conduct argumentative analysis of these debates, specifically identifying argument components (premises and claims) within the debate transcripts. Thus, this first version of DispuTool offered three main features:

1. **Exploration** of arguments in presidential debates from 1960 to 2016
2. Additional **exploratory functions**, such as Named Entity Recognition
3. **Argumentative analysis** of political debate transcripts not included in the original dataset

A detailed explanation of the first version of these features is provided in Section 5.1.1, 5.1.2, and 5.1.3. These features served as the foundational elements for the last version of DispuTool (2.0), which is comprehensively detailed in Section 5.2.

5.1.1 Exploration of U.S. Presidential Debates

DispuTool 1.0 enhances its analytical capabilities by employing Named Entity Recognition, utilizing the Stanford Named Entity Recognizer[2]. This feature allows users to explore the debate corpus through a specialized filtering system based on various entity categories. Researchers can refine their analysis by focusing on specific types of named entities, including persons, locations, nationalities, organizations, and religions. The tool also enables temporal analysis by allowing users to filter debates based on the year they occurred, spanning from 1960 to 2016. Furthermore, DispuTool 1.0 facilitates speaker-specific analysis by providing a comprehensive list of all debate candidates, allowing users to isolate and examine the contributions of individual speakers, as shown in Figure 5.1.

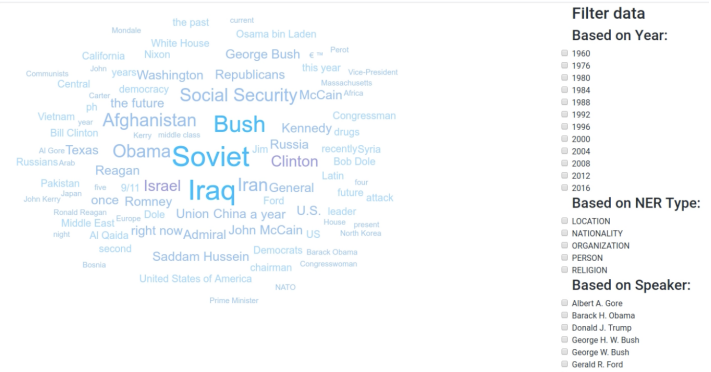


Figure 5.1: Named Entity Recognition section.

This extensive approach to data exploration empowers researchers to conduct extensive investigations into the use of named entities in political discourse, potentially revealing patterns or topic focus across different debates, years, or speakers.

5.1.2 Argumentative Analysis

DispuTool 1.0 provides researchers with a sophisticated interface to analyze a corpus of 41 annotated U.S. presidential debates. This corpus has been meticulously coded to identify argumentative components, specifically premises and claims. The tool's primary exploration feature allows users to:

- Select a specific debate of interest from the corpus.
- View a comprehensive list of premises and claims presented in the chosen debate.
- Identify the candidate who proposed each argumentative component.
- Note the date on which the debate occurred.

[2] https://nlp.stanford.edu/software/CRF-NER.html

DispuTool 1.0 offers a visual representation of the entire debate transcript. Users can opt to highlight premises and claims using distinct colors, allowing for immediate visual identification of argumentative structures within the text. This feature is particularly useful for identifying patterns in argumentation or comparing argumentative strategies between candidates.

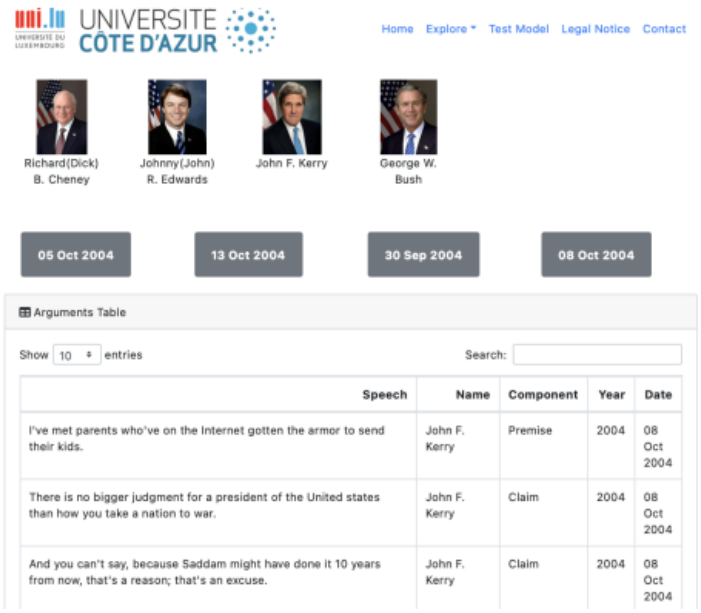


Figure 5.2: Exploration of argument components within a specific debates.

Figure 5.2 illustrates the interface for exploring argument components in a specific debate, while Figure 5.3 demonstrates the transcript visualization feature with color-coded highlighting of argumentative elements.

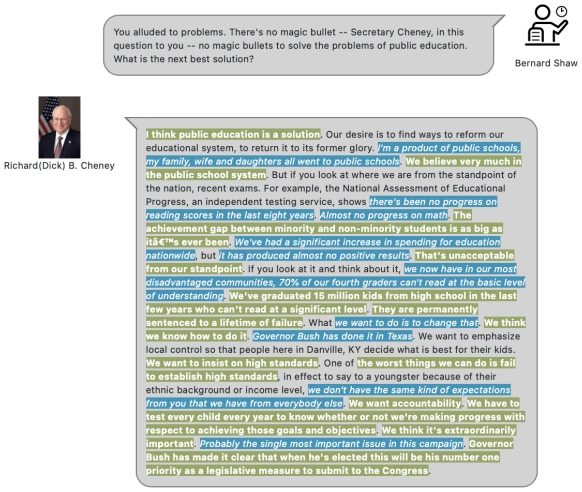


Figure 5.3: Visualizing argument components in the debate context.

This functionality tool enables researchers to conduct in-depth analyses of political debates, examining both the content and structure of arguments presented in these challenging debates.

5.1.3 Analyze Your Debate

DispuTool 1.0 extends its analytical capabilities beyond the pre-existing corpus by offering a dynamic text analysis feature. Users can input any debate transcript of interest into a dedicated text field, enabling *ad hoc* argumentative analysis. Once submitted the text, the tool processes the input using advanced NLP techniques to identify key argumentative components. The system then presents the analyzed text with visual enhancements: claims are highlighted in green, while premises are accentuated in blue. This color-coded output provides researchers with an immediate visual representation of the argumentative structure within the text, facilitating rapid identification of key arguments and their supporting evidence, as shown in Figure 5.4. This feature significantly broadens the tool's applicability, allowing researchers to apply DispuTool 1.0's analytical framework to a wide range of political discourses beyond the initial corpus, thus enhancing its value for comparative studies and contemporary debate analysis.

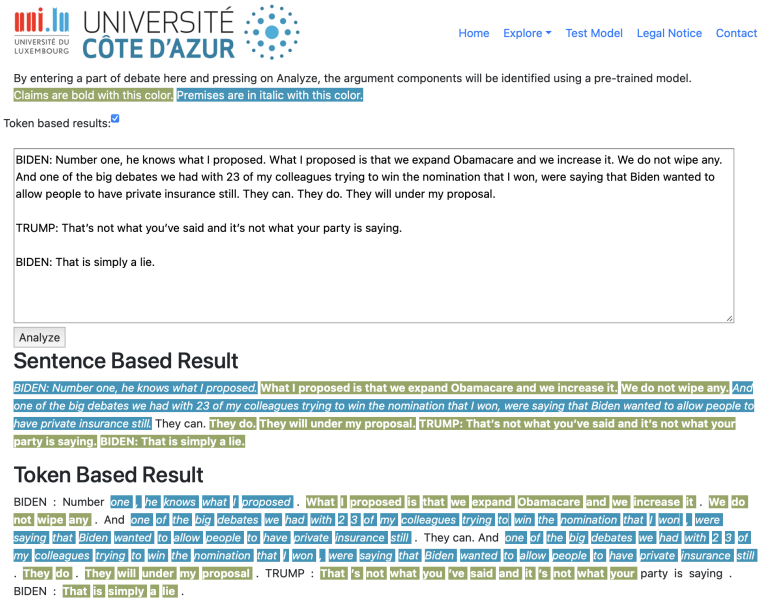


Figure 5.4: Visualizing argument components in the debate context.

5.1.4 Experimental Setting and Results

A pipeline structure was developed [62] for the identification of argument components. This approach comprises two sequential tasks: *1)* distinguishing argumentative sentences from non-argumentative parts to establish argument boundaries, *2)* classifying the identified components as either premises or claims. Three distinct classifiers were trained for each step of the pipeline. The first classifier employs a Support Vector

Machine (SVM) with structural (e.g., sentence length), semantic (e.g., sentiment polarity, named entities), and linguistic features (e.g., words, part of speech). The second consists of a bidirectional Long Short-Term Memory (LSTM) neural network with pretrained word embeddings. The third implements a Feed-Forward neural network using the same feature set as the SVM classifier. Performance metrics (precision *p*, recall *r*, and *F1 Score*) for Task 1 classification using SVM with RBF Kernel and all features yielded *p* 0.851, *r* 0.853, and *F1* 0.823. For Task 2, the LSTM network with word embeddings achieved *p* 0.673, *r* 0.673, and *F1* 0.673. Based on these results, the LSTM model with word embedding features, demonstrating the highest performance, was integrated into DISPUTool 1.0 for argument component identification.

5.2 DispuTool 2.0

A major contribution of this thesis is the upgrade of DispuTool's architecture and functionality, necessitated by the rapid evolution of underlying technologies. The second version of the tool [49] incorporates two significant improvements:

1. **Model Upgrade**: The core model that supports the "Analyze Your Debate" feature (see Section 5.1.3) had a major upgrade. This refinement aimed to improve the model's performance in political debate analysis, with the aim to detect argument components and also predict argument relations between them.
2. **Expanded Analytical Capabilities**: The debate exploration functionality (see Section 5.1.1) was augmented to include a systematic analysis of argument fallacies employed in political debate. This addition facilitates a more nuanced understanding of argumentative structures and persuasive techniques utilized in debates.

These advancements focus on incorporating diverse analytical perspectives to examine debates, thereby facilitating a more comprehensive and nuanced understanding of complex argumentative structures. Moreover, the forthcoming implementation of the state-of-the-art model detailed in Chapter 4 is expected to significantly enhance the tool's analytical capabilities. This upgrade, coupled with the introduction of novel visualization techniques and the dataset update with the 2020 debates between Biden and Trump, will enable a more granular analysis of debates and rhetorical fallacies. These ongoing developments underscore the dynamic nature of DispuTool 2.0, reflecting its continuous evolution in response to the intricate demands of contemporary political discourse analysis. The iterative refinement process ensures that the tool remains at the forefront of computational approaches to debate evaluation and rhetorical analysis.

5.2.1 Argumentative Analysis

In DispuTool 2.0, it is possible to explore the corpus made of 44 U.S. presidential debates annotated with argumentative components and relations. When the user selects one of the debates, three argumentative elements are shown:

1. The argumentative components are highlighted in the textual arguments put forward by each candidate, and a label 'claim' or 'premise' is associated with these pieces of text (as in DISPUTool 1.0)
2. The relations holding between the identified components are identified and labeled to indicate whether it is a support or an attack relation
3. Fallacious arguments are highlighted in the text and associated with one of the following 6 classes of fallacies: Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, and Slogan.

This improved visualization is shown in Figure 5.5.

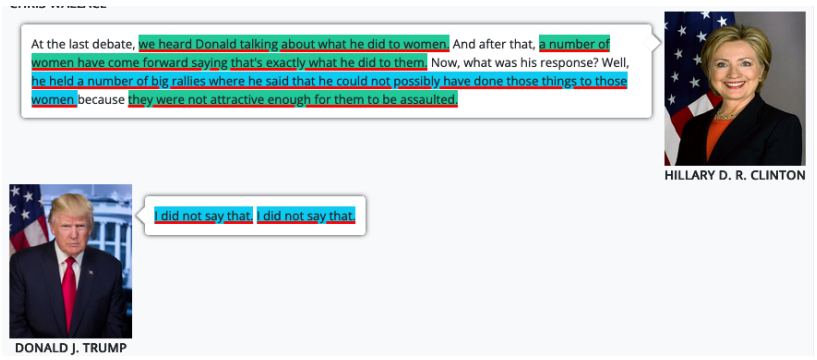


Figure 5.5: Exploration of debates with fallacies highlighted in red.

In addition, DispuTool 2.0 now incorporates a graph-based visualization feature for political debates. This feature represents argument components (claims or premises) as nodes and their relations (supporting or attacking) as edges, as showed in Figure 5.6. This graphical approach offers an alternative analytical perspective, enabling rapid identification of particular debate points and argumentative structures. By transforming verbal exchanges into visual data, users can more efficiently discern patterns and interconnections within the discourse. This visualization technique complements traditional textual analysis, aligning with contemporary trends in computational argumentation and potentially opening new opportunities for quantitative and qualitative research in political communication studies.

5.2.2 Exploration of U.S. Presidential Debates

This section delineates the enhancements implemented in the exploration module of political debates within DispuTool 2.0. The Named Entity Recognition (NER) visualization has gone through significant evolution, incorporating multiple novel representational techniques to support accurate data analysis:

- **Word Cloud Visualization** (Figure 5.7): This representation provides a frequency-based depiction of identified entities, allowing for rapid assessment of prominent topics and candidates within the debates.

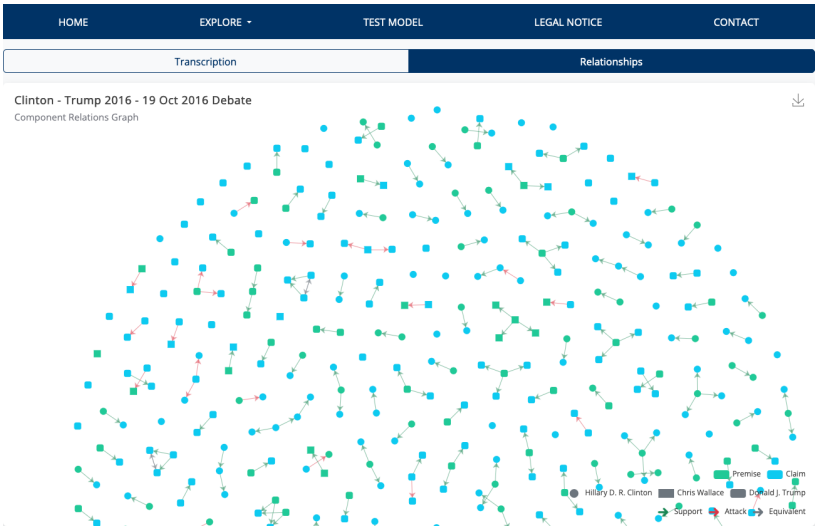


Figure 5.6: Exploration of debates as a graph.

- **Bubble Chart** (Figure 5.8): Similar to the word cloud, this visualization represents entities as bubbles.
- **Sankey Diagram** (Figure 5.9 and Figure 5.10): This flow diagram illustrates the distribution and interconnections of topics and fallacies addressed by candidates, facilitating the analysis of thematic patterns and shifts across debates.
- **Stacked Area Chart** (Figure 5.11): This time-series visualization presents an alternative perspective on topic coverage, allowing for the examination of thematic evolution over multiple years.

These varieties of visualization techniques collectively enhance the analytical capabilities of DispuTool 2.0, providing researchers with a holistic approach to exploring and interpreting political debate data. By offering various visualization modalities, the tool facilitates both macro-level trend analysis and micro-level examination of specific debate elements.

5.2.3 Analyze Your Debate

The latest version of DispuTool (2.0) offers advanced functionality for argumentation analysis, extending beyond mere identification of argumentative components to predict potential inter-component relationships within user-provided political debate texts. The augmented analytical output is presented through two complementary visualization formats:

1. **Annotated Text Display**: The input text is processed to highlight identified argumentative components, differentiating between claims and premises through distinct visual cues.

Figure 5.7: New NER visualization of DispuTool 2.0.

2. **Relational Graph Representation**: A graph diagram is generated, wherein nodes represent argumentative components (claims and premises), and edges denote the predicted relationships (support or attack) between these components.

Figure 5.12 provides an illustrative example of this dual-visualization approach, showing the tool's capacity for comprehensive argumentation analysis.

The pipeline architecture involved in the component identification and relation prediction tasks represents a significant advancement over the model employed in the previous version (detailed in Section 5.1.3). This updated architecture enhances the tool's analytical capabilities and efficiency. However, it is noteworthy that, as of this writing, efforts are in progress to further upgrade the system to incorporate the novel architecture delineated in Chapter 4. This upcoming version leverages the latest libraries and promises superior performance metrics compared to the current implementation. This ongoing development underscores DispuTool 2.0's dedication to continuous improvement and adaptation to emerging technologies in the field of computational argumentation analysis.

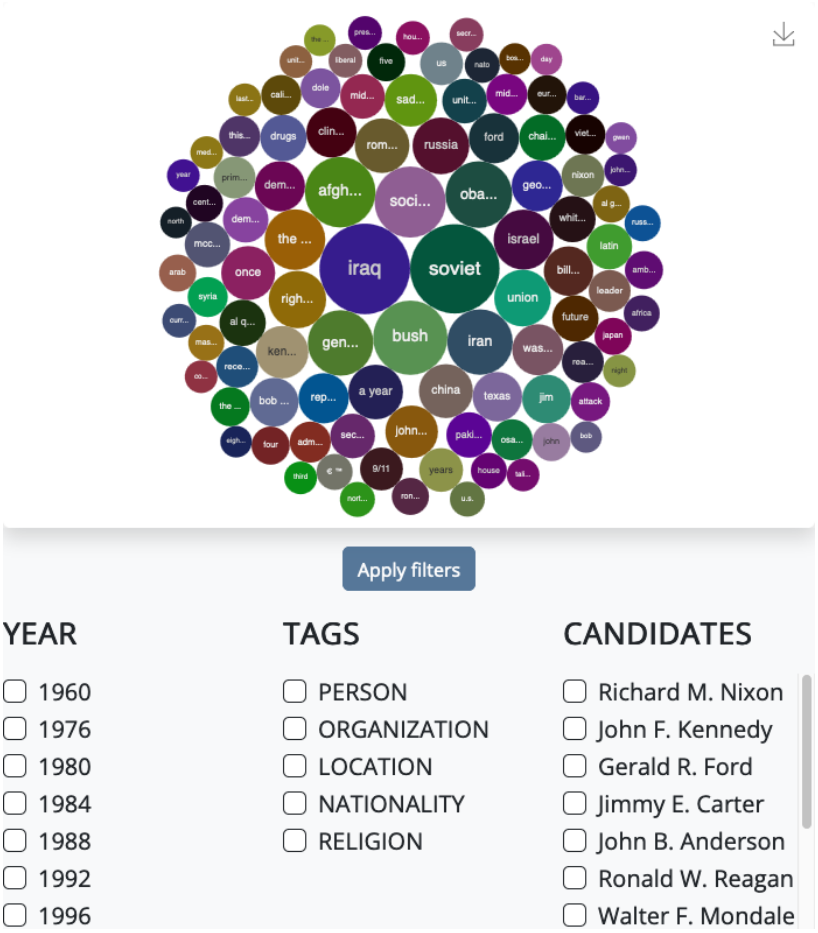


Figure 5.8: New Bubble Chart visualization of NER entities.

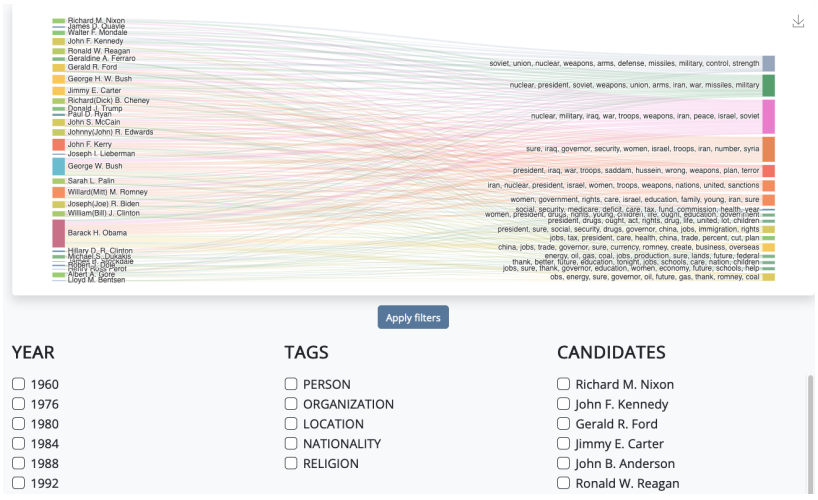


Figure 5.9: Visualization of topics in debates using a Sankey Diagram.

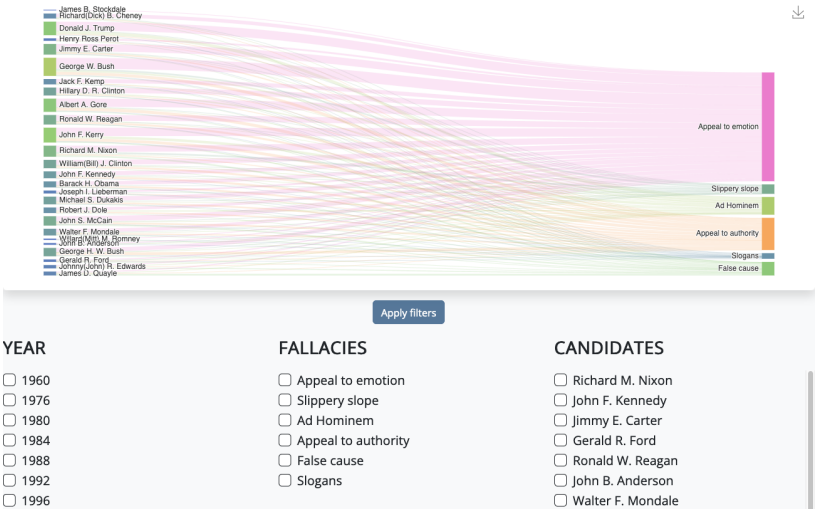


Figure 5.10: Visualization of fallacies in debates using a Sankey Diagram.

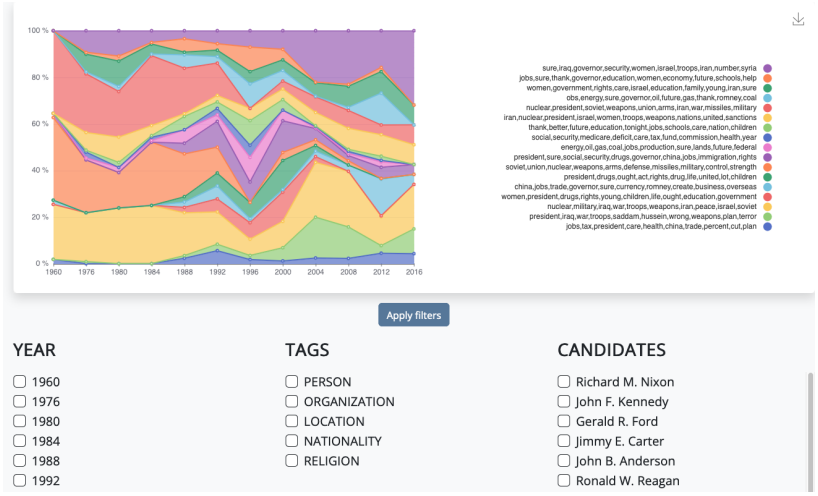


Figure 5.11: Visualization of covered topics in years using a Stacked Area.

5.2.4 Experimental Setting and Results

The pipeline architecture employed in this version of DispuTool adopts the same concepts detailed in Section 4.1.2 and 4.2, based on Mayer, Cabrio, and Villata [110] approach. For the component identification task, the initial representation is computed using the BERT base model [36], which is fine-tuned over 15 epochs using an Adam optimizer with a learning rate of 6e-5 and a maximum sentence length of 64 tokens. Subsequently, the sentence representation goes into sequential processing through a Gated Recurrent Unit (GRU) [26] followed by a Conditional Random Field (CRF) [88]. This configuration achieved an F1 Score of 0.79 on the test set for component identification.The relation prediction task employs a sequence classification approach to jointly model relations between argumentative components. This task utilizes a bidirectional transformer architecture for classifying all component combinations. A linear layer

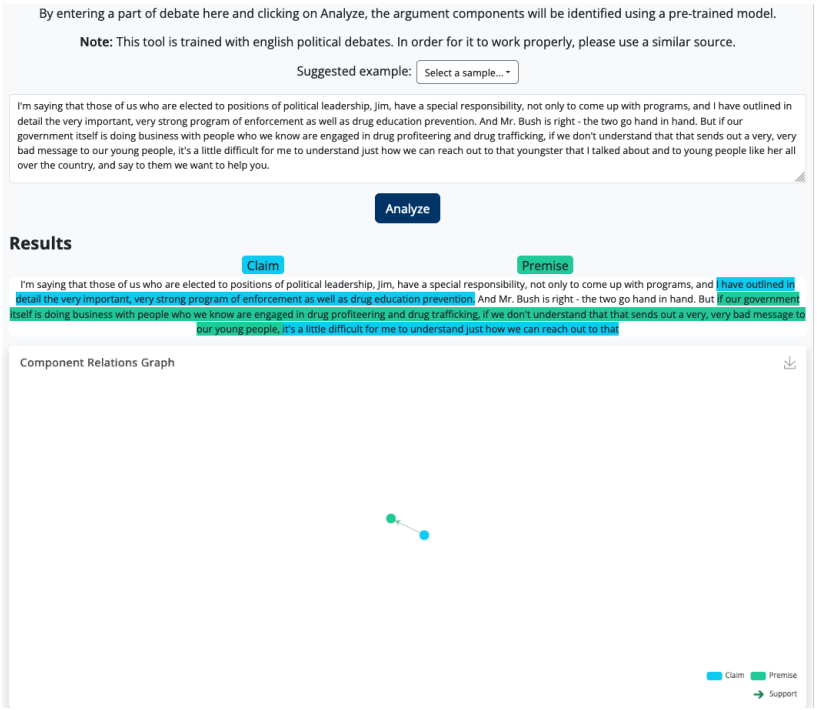


Figure 5.12: New visualization of the analyzed debate.

with a softmax function is applied for three-way classification, categorizing relations as Support, Attack, or NoRelation. The base model used is RoBERTa [103] with pre-trained weights, which is then fine-tuned with a learning rate of 6e-5, a batch size of 8, and a maximum sentence length of 64 sub-word tokens per input example over 15 epochs. This configuration for relation prediction attained a macro F1 Score of 0.60 on the test set.

5.3 Summary

This chapter presents in detail DISPUTool 2.0, an advanced web-based tool for multi-layer argumentative analysis of political debates. Building upon its previous version, this version introduces new capabilities in argument mining, specifically designed for English political debates. The tool allows users to explore and analyze US presidential debate transcripts from 1960 to 2020, automatically identifying argumentative components (premises and claims), and their relations (support and attack), and examine fallacious arguments within the debates. DISPUTool 2.0 employs state-of-the-art natural language processing techniques, including transformer-based models like BERT and RoBERTa, to achieve high accuracy in argument component detection (F1 Score of 0.79) and relation prediction (macro F1 Score of 0.60). Notable features include the ability for users to analyze their own political debate texts, and a named entity recognition function.

Chapter 6

Fallacy Identification

This chapter explores two related tasks in the domain of argumentative analysis within political debates: the multi-class classification of argumentative fallacies and their subsequent detection and classification in debate contexts. The initial task involves transformer-based models designed to handle a large amount of text to classify fallacies at both coarse and fine-grained levels. This work is based on the ElecDeb60to16 dataset. The second task addresses the more nuanced challenge of fallacy detection in the larger context of debates. This advanced approach, developed using the updated ElecDeb60to20 dataset, focuses on coarse-grained annotations and employs architectures specifically optimized for token classification. A key innovation in both approaches is the integration of argumentative and textual features, which has yielded substantial improvements in model performance.

The detection and classification of argumentative fallacies in political debates represent crucial tasks in the field of computational argumentation analysis. These goals are driven by the pressing need to enhance public discourse and democratic processes. In an era of information overload and increasing political polarization, the ability to automatically identify logical flaws and manipulative rhetorical techniques in political arguments is crucial. This capability not only aids in creating a more informed society but also contributes to the development of critical thinking skills essential for navigating complex political landscapes. The distinction among different types of fallacies is crucial in political discourse analysis, as various fallacies can have different impacts on the reasoning process and public opinion. In the pragma-dialectical theory of argumentation, fallacies are conceptualized as "derailments of strategic maneuvering" that violate the rules of rational argumentative discussion [42, 41]. These derailments are particularly significant in political discourse, where informal fallacies are strategically employed by politicians to advance their positions [174, 96]. This chapter presents a comprehensive approach to analyzing fallacies in political debates, progressing from classification to more advanced detection and classification tasks. The thesis evolved to include two main tasks: firstly, fine- and coarse-grained fallacy classification, and secondly, the identification and classification of fallacies within political debates. This

two-stage approach allowed us to first establish a solid foundation in fallacy categorization before tackling the more complex challenge of detecting fallacies within the broader context of debates. The approach starts with proposing an annotation scheme for fallacies in political debates, enabling more fine-grained classification and identification tasks. This initial step was crucial in laying the groundwork for deeper analysis.

To address the challenges of fallacy analysis in political debates, I propose a new methodology inspired by [160]. This approach allows the employed models to take a broader point of view while improving their performance in classifying and identifying fallacies. The methodology incorporates additional contextual and engineered information, enabling a more nuanced understanding of the argumentative structures in political discourse. The research presented in this chapter addresses these critical needs through the aforementioned interrelated tasks: multi-class classification of argumentative fallacies and their contextual detection within political debates. These tasks build upon recent pragmatic approaches that have redefined fallacies as infringements of argumentative engagement rules [43] and as inappropriate dialectical shifts across dialogue types [165]. For the two tasks, supervised approaches are proposed and tested on the ElecDeb60to16 and ElecDeb60to20 datasets, respectively. These extensive temporal ranges allow for a comprehensive analysis of fallacy usage in American political discourse over decades. Section 6.1 describes the first approach: the fallacy classification task applied to U.S. political debates from 1960 to 2016. It details the rationale behind the choice of fallacy categories, focusing on those more inherent and widely used in political debates. Section 6.2 extends the research by addressing a more complex challenge: the identification and classification of fallacies in political debates, including those of 2020. This analysis builds upon the insights gained from previous sections. This progression demonstrates our commitment to developing increasingly sophisticated analytical tools, extending the analysis to more recent political debates between Biden and Trump. Both sections include subsections designed to present the experimental setup, discuss the results, and conclude the error analysis.

6.1 Fallacy Classification

This work has been realized with the first version of ElecDeb60to16 dataset (see Section 3.2.3), so the fallacies are annotated up to the 2016 Trump-Clinton debate. In addition, experiments were performed on both coarse-grained and fine-grained annotations on a total of 1,628 argumentative fallacies. In this work, I tackle the challenge of automatically classifying fallacious arguments in political debates. The primary objective is to create a system capable of automatically recognizing and categorizing different types of fallacious arguments commonly used in political debates. To achieve this, it has been necessary to establish and annotate the most prevalent categories of fallacious arguments found in political discourse. The categories taken into consideration are Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, and Slogan. Thus, the task was approached as a sequence classification problem.

The source code and associated materials for this work are publicly available in the GitHub repository at https://github.com/pierpaologoffredo/IJCAI2022.

6.1.1 Experimental Setup

As already mentioned in Section 6.1, to address the fallacy classification task, I addressed it as a sequence classification problem. The methodology consisted of two main stages:

1. **Multi-class Classification**: Initially, a classifier to categorize the various fallacies identified in the debates was developed.
2. **Feature Enhancement**: Subsequently, the classifier has been refined by incorporating argumentation-based features. Specifically, information about argument components and their argument relations within each fallacious argument were integrated.

This two-step approach allowed to first establish a baseline classification model and then enhance its performance by leveraging the structural elements of argumentation present in the debates.

Initially, I considered BERT [36] and RoBERTa [103] as baseline models. However, the specific characteristics of presidential debate data presented significant challenges. The extensive length of debate transcripts exceeds the typical input limitations of classical transformer models, necessitating a more sophisticated approach. BERT, for instance, has a maximum sequence length of 512 tokens. While this is generally sufficient for individual fallacious arguments, it falls short when incorporating the full context of a debate segment. Each speech turn often contains more than 512 tokens, especially when the candidate is able to answer without interruptions. To address these limitations, I considered more advanced Transformer models specifically designed to handle larger inputs. Two models stood out as particularly suitable for my task:

- **Longformer** [13]: This model employs an attention mechanism that scales linearly with sequence length, enabling it to process documents of thousands of tokens or more. Longformer's architecture combines windowed local-context self-attention with global attention, allowing it to build both detailed contextual representations and comprehensive sequence representations. Like RoBERTa, Longformer is pre-trained using the Masked Language Modeling (MLM) approach.
- **Transformer-XL** [35]: This model excels at learning dependencies beyond fixed dimensions without affecting temporal coherence. By integrating recurrence and relative positional encoding, Transformer-XL can model longer-term dependencies more effectively than traditional RNNs and vanilla Transformers.

In my experimental design, I incorporated a range of features[1] to enhance the fallacious argument classification task, aligning with established frameworks in computational argumentation [20, 90]. These features can be categorized into three main types:

- **Political discourse context**: This refers to the global speech turn in which the fallacious argument is placed.
- **Argument component labels**: These are derived from the fundamental building blocks of argumentation, namely: claims and premises.
- **Argument relation labels**: These describe the relations between argument components, including: attack, support, and equivalent.

In my initial experiments, I compared several models to classify the main categories of argumentative fallacies in political debates. I used the Hugging Face Transformers library (version 1.7.0) with PyTorch for implementation. All models shared these hyperparameters: learning rate of 5e-5, dropout rate of 0.1, 3 training epochs, and Adam optimizer. Model-specific settings were as follows. BERT and RoBERTa used a batch size of 8 and a maximum sequence length of 128 tokens for the fallacious argument snippet. Transformer-XL required a batch size of 1 due to high memory demands. It processed fallacious argument snippets with a max length of 128 tokens and context speech up to 8,192 tokens. Longformer used a batch size of 8, with max lengths of 128 tokens for fallacious argument snippets and 4,096 tokens for speech context. The varying sequence lengths reflect each model's capacity for handling long-range dependencies. BERT and RoBERTa could only process the fallacious argument snippets in full, while Transformer-XL and Longformer were able to handle a larger speech turn context with the fallacy snippet. I reduced Transformer-XL's batch size to address its high memory usage, allowing it to process longer sequences despite resource constraints.

Once found the model with the best performance, I performed additional experiments to *1)* classify the 14 fallacious argument sub-categories, i.e., General Ad Hominem, Bias Ad Hominem, Tu Quoque, Name-calling Labeling, Appeal to Fear, Appeal to pity, Loaded Language, Flag Waving, Without Evidence, False Authority, Popular Opinion, False Cause, Slippery Slope, and Slogan; and *2)* classify the main categories, enriching the dataset with the argument component and relation labels in an **ablation test setting**.

For my ablation study, I adapted the methodology from Vorakitphan et al. [160] to test feature combinations. I built on the best-performing model (Longformer), adding argumentation features and debate context. Each training sample has four components: *1)* dialogue context, *2)* fallacious argument snippet, *3)* argument component label, and *4)* argument relation label. These are processed through a Transformer Model to create embedded vectors, each fed into a transformer-based classifier to produce a logit. Figure 6.1 shows this architecture. This method lets me evaluate each feature's impor-

[1] I extracted the argumentative features from the pre-existing annotations in the ElecDeb60To16 dataset [63].

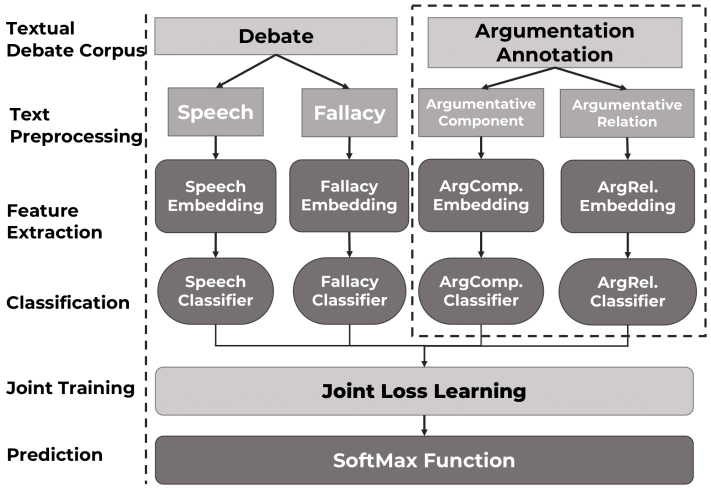


Figure 6.1: Architecture employed for the fallacy classification task. Figure drawn from [51].

tance in fallacy classification, understanding their individual and combined effects on model performance. Adam optimizer, 0.1 dropout, and Cross-Entropy loss function are employed. I calculate loss separately for each component: fallacy snippet ($loss_snippet$), speech context ($loss_speech$), argument component ($loss_{ArgComp}$), and argument relation ($loss_{ArgRel}$). I combined these losses using a joint learning approach:

$$loss_{joint_loss} = \alpha * \frac{(loss_{speech} + loss_{snippet} + loss_{ArgComp} + loss_{ArgRel})}{N_{loss}} \tag{6.1}$$

Here, α is 0.5, and N_{loss} is the number of loss elements. This joint loss is used for backpropagation, allowing the model to learn from all features simultaneously.

The dataset was partitioned into training and test sets using the train_test_split function from the scikit-learn library [127]. The split was performed with a ratio of 80% for the training set and 20% for the test set. To ensure that the distribution of fallacy types was consistently represented in both sets, stratification was applied using the fallacy labels column. This stratified sampling approach maintains the proportion of each fallacy category in both the training and test sets, mitigating potential biases that could arise from imbalanced class distributions.

6.1.2 Result and Discussion

Table 6.1 presents a general overview of experimental results, including both initial experiments and ablation tests. The performance metrics considered are precision, recall, and F1 Score, providing a holistic evaluation of each approach. Analysis of the various models reveals a clear pattern: the Longformer model, coupled with

the $loss_{joint_loss}$ method, consistently outperforms other approaches, achieving the highest F1 Score in the 6-category classification task. This superiority likely comes from Longformer's ability to handle longer sequences, a crucial advantage when analyzing extended political debates. The model's performance is further enhanced when argumentation features are incorporated, demonstrating significant improvement over existing techniques such as those proposed by [58, 57]. The introduction of argumentation features provides a turning point in this study. By incorporating component labels, relation labels, or both, it is possible to observe a considerable improvement in model performance, with the combination of both features yielding the highest F1 Score of 0.84. This underscores the critical role of argument structure in accurate fallacy detection, aligning with established theories in argumentation studies. However, these findings also highlight challenges in fine-grained fallacy classification. When increasing fallacy categories from 6 to 14, a significant performance drop is shown, with the F1 Score falling from 0.61 to 0.42. This drop suggests that classifying between more nuanced fallacy types poses a greater challenge, possibly due to increased complexity and potential overlap between subcategories. The promising results of the Longformer model led me to adopt it as the proposed architecture for further investigation. Consequently, I extended the experiments to tackle the more granular task of classifying fallacious arguments into their respective sub-categories.

Model	**Fallacy Cat.**	**Arg. Feat.**	**Prec.**	**Rec.**	**F1 Score**
BERT	6	–	0.62	0.55	0.55
RoBERTa	6	–	0.58	0.56	0.53
Longformer	6	–	0.64	0.6	0.57
Longformer $loss_{joint_loss}$	6	–	0.66	0.61	**0.61**
Transformer-XL	6	–	0.61	0.45	0.47
Transformer-XL $loss_{joint_loss}$	6	–	0.61	0.51	0.53
Longformer $loss_{joint_loss}$	14	–	0.44	0.45	0.42
Longformer $loss_{joint_loss}$	6	Comp. Label	0.88	0.81	0.83
Longformer $loss_{joint_loss}$	6	Rel. Label	0.87	0.81	0.83
Longformer $loss_{joint_loss}$	6	Comp. & Rel. Label	0.84	0.85	**0.84**

Table 6.1: Results of the baseline and ablation experiments.

Table 6.2 offers a detailed overview of the model's performance on this task. In particular, the model exhibits strong performance across several specific fallacy types, particularly excelling in identifying instances of Flag Waving, Slogans, Loaded Language, and arguments Without Evidence. This success is not entirely unexpected, given that these categories are the most prevalent in our dataset, providing the model with ample examples for learning.

The ablation test results, presented in Table 6.3, clearly demonstrate the significant impact of argumentative features on fallacy classification performance. Across all main

	Prec.	**Rec.**	**F1 Score**
Ad Hominem	0.47	0.60	0.52
Appeal to Fear	0.47	0.41	0.43
Appeal to Pity	0.60	0.47	0.51
Popular Opinion	0.68	0.49	0.50
Circumstantial Ad hominem	0.27	0.30	0.28
False Authority	0.00	0.00	0.00
False Cause	0.19	0.44	0.27
Flag Waving	0.62	0.70	**0.65**
Loaded Language	0.85	0.82	**0.83**
Name-Calling, Labeling	0.33	0.22	0.27
Slippery Slope	0.45	0.31	0.32
Slogan	0.69	0.69	**0.68**
Tu Quoque	0.00	0.00	0.00
Without Evidence	0.48	0.78	**0.57**
Accuracy			0.63
Macro avg	0.44	0.45	0.42
Weighted avg	0.62	0.63	0.61

Table 6.2: Results of fallacy classification considering the subcategories.

fallacy categories, the inclusion of argument components and relation labels substantially improves the model's performance. Context alone (ctx) provides a baseline performance, but the addition of argumentative features consistently boosts results. For instance, Ad Hominem detection improves from 0.56 to 0.81 when both component and relation features are included. Similarly, Appeal to Authority scores a remarkable increase from 0.65 to 0.91. Notably, different fallacy types benefit differently from various feature combinations. Slogans, for example, show the highest improvement with relation labels (0.88), while Appeal to Emotion benefits most from a combination of all features (0.94). The macro average F1 Score increases from 0.61 with context alone to 0.84 when all argumentative features are included, emphasizing the crucial role of argument structure in fallacy detection. This improvement highlights how argumentative features provide essential context beyond the raw speech content, enabling more nuanced and accurate classification of fallacious arguments.

6.1.3 Error Analysis

Table 6.4 shows examples of misclassified fallacies, revealing challenges in the model's categorization. Many errors involve confusion between Slogans, Appeals to Emotion, and Ad Hominem arguments. These fallacy types often use similar language to manipulate audience sentiment, making distinction difficult. The phrase "It is time for a change" was misclassified as an Appeal to Emotion instead of a Slogan, while "We have to change the culture of America" was incorrectly labeled as a Slogan rather than an Appeal to Emotion. This highlights the fine line between emotional appeals and persuasive phrases in political discourse. The model sometimes misclassifies complex fallacies as simpler, emotion-based ones. A Slippery Slope argument about taxes was

	ctx	**ctx + comp**	**ctx + rel**	**ctx + comp + rel**
Ad Hominem	0.56	0.85	0.81	0.81
Appeal to Authority	0.65	0.85	0.84	0.91
Appeal to Emotion	0.85	0.93	0.93	0.94
False Cause	0.43	0.80	0.82	0.80
Slippery Slope	0.50	0.78	0.79	0.84
Slogans	0.67	0.76	0.88	0.77
Accuracy	0.75	0.88	0.89	0.89
Macro avg	0.61	0.83	0.83	**0.84**
Weighted avg	0.74	0.88	0.89	0.89

Table 6.3: Ablation test results considering the additional features: argument component label and argument relation label. All the results refer to the macro average F1 Score metric.

misinterpreted as an Appeal to Emotion, and a False Cause argument about foreign policy was labeled as Ad Hominem. This suggests the model may be overly sensitive to emotional language, missing logical structures in certain fallacy types. Ad Hominem arguments are frequently misclassified as Appeals to Emotion, as seen in the last three examples. The model struggles to differentiate personal attacks from general emotional manipulation, especially in provocative language.

Fallacious snippet	**True**	**Predicted**
It is time for a change.	Slogan	App. to Emotion
We have to change the culture of America.	App. to Emotion	Slogans
I think if you raise taxes during a recession, you head to depression.	App. to Emotion	Slippery Slope
Bill Clinton, as President, has provided that kind of leadership. We are more secure and stronger today because of Bill Clinton's handling of foreign policy.	False Cause	Ad Hominem
an old washed-up terrorist	Ad Hominem	App. to Emotion
I do want to bring up the fact that you were the one that brought up the words super-predator about young black youth.	Ad Hominem	App. to Emotion
the most ruthless, fanatical... leaders that the world has ever seen	Ad Hominem	App. to Emotion

Table 6.4: Some examples of misclassified snippet using the best model.

To summarize, these findings underscore a broader challenge in the field of argumentation. Fallacies remain a controversial issue, particularly when applying theoretical frameworks to real-world scenarios like political debates. As noted by Boudry et al. [18], the traditional argumentation schemes used to identify invalid or flawed reasoning often fall short when confronted with the complexities of actual discourse. The core of this problem lies in the tendency of these schemes to abstract away from the specific content and dialectical context of the fallacy. This abstraction, while useful for

theoretical analysis, can lead to oversimplification when applied to the nuanced and often ambiguous arguments found in political debates. The proposed model's struggles with certain fallacy types reflect this fundamental challenge. The difficulty in distinguishing between closely related fallacies, or in recognizing complex fallacies within emotionally charged language, mirrors the broader issues faced by argumentation theorists and practitioners.

Moving beyond the task of categorizing pre-identified fallacious arguments, I now turn the attention to the more complex problem of fallacy detection and classification within the wider context of political debates. This evolution requires not only determining the category of a fallacy but also detect its exact textual boundaries within larger discourse segments, significantly increasing the task's complexity and real-world applicability. This new, more intricate task of fallacy detection and classification, and its associated challenges, will be thoroughly explored in Section 6.2 of this thesis.

6.2 Fallacy Detection & Classification

This section extends the investigation beyond the multi-class classification of fallacies discussed previously, delving into the more complex task of detecting and classifying fallacies within the broader context of political debates. This stage represents a natural evolution in my thesis, significantly expanding the scope and applicability of Argument Mining techniques (see Section 2.2) to fallacies in political debates. While the preceding chapter focused on classifying pre-identified fallacies, this section addresses the dual challenge of both identifying and categorizing fallacies contained within extended political discourse. This advancement not only enriches the field of Argument Mining but also provides valuable tools for analyzing and understanding political debates. My approach builds upon established methods for identifying and classifying argumentative components, adapting and extending these techniques to address the unique challenges posed by fallacious reasoning in political contexts. By developing and evaluating methodologies capable of automatically detecting and categorizing fallacies within political debates, I aim to contribute innovative insights and practical tools to the field of computational argumentation analysis. While significant progress has been made in the field of fallacy analysis, existing approaches have primarily focused on the classification of pre-identified fallacious text snippets across a finite set of labels [56, 57, 51, 157, 5, 161, 139]. This emphasis has left a critical gap in the research: the challenging task of identifying fallacious text snippets and determining their boundaries within broader discourse remains under-investigated. This research addresses this gap and makes two primary contributions to the field. First, the ElecDeb60to16-fallacy [51] dataset has been expanded incorporating debates from the 2020 presidential campaign (Trump-Biden), complete with argument (component and relation) and fallacy annotations, as detailed in Section 3.2.3. This extension provides a more current

and robust resource for fallacy analysis in political discourse. Second, I propose an innovative method for detecting fallacious text snippets within political debates and classifying them according to the six established fallacy categories. The approach leverages state-of-the-art Transformer models, integrating both argument-based and engineered features to effectively identify and categorize fallacious arguments.

The source code and associated materials for this work are publicly available in the GitHub repository at `https://github.com/pierpaologoffredo/FallacyDetection`.

The method of how fallacies are detected and classified with the specifications of the experimental setup is described in Section 6.2.1. Subsequently, in Section 6.2.2 the results are presented, and observed problems are discussed in Section 6.2.3.

6.2.1 Experimental Setup

Expanding the analysis of fallacy distribution and characteristics in political debates, I approach the task of fallacy detection as an information extraction problem. The objective is to identify and classify textual segments within the debates that correspond to the six previously annotated fallacy categories. To accomplish this, I employ the BIO (Begin, Inside, Outside) tagging scheme, a standard format in natural language processing for sequence labeling tasks. Specifically, we utilize the following tags: B-AdHominem, I-AdHominem, B-AppealtoAuthority, I-AppealtoAuthority, B-AppealtoEmotion, I-AppealtoEmotion, B-FalseCause, I-FalseCause, B-SlipperySlope, I-SlipperySlope, B-Slogans, I-Slogans, and O. This results in a thirteen-class classification problem, where each token in the debate transcript is assigned one of these predefined labels. To enhance the contextual understanding of fallacies, I extend the framework beyond isolated sentences. The framework considers not only the sentence containing the potential fallacy but also its immediate textual structure – the preceding and following sentences. This triadic structure provides a richer representational context for fallacy detection. In cases where the fallacious sentence occurs at the beginning or end of a dialogue, I adopt this approach by excluding the preceding or following sentence, respectively.

To address the challenges of fallacy detection and classification, this research employed transformer-based architectures, leveraging their advanced capabilities in natural language processing tasks. This approach employs these models in two configurations: a basic setup and a specialized configuration optimized for token classification[2]. This methodology is motivated by recent advancements in fallacy detection and classification research [32, 160, 51], which have demonstrated the efficacy of augmenting transformer-generated text representations with non-textual features. In this implementation, I further enhance the specialized architecture by incorporating additional argumentative and textual features (i.e., argument component and relation labels, and PoS tags). This hybrid approach allows capturing both the nuanced linguistic patterns characteristic of fallacious arguments and the broader argumentative structures

[2] `https://huggingface.co/docs/transformers/tasks/token_classification`

within which they occur. By combining state-of-the-art natural language processing techniques with domain-specific argumentative features, my method aims to achieve a more robust and contextually aware system for fallacy detection and classification in political discourse. This approach not only builds upon established research in the field but also pushes the boundaries of what's possible in automated analysis of complex argumentative structures.

For this task, experiments are conducted with some baseline models and some pre-trained transformer models, i.e., BERT [36] with LSTM [70] and BiLSTM [142], DeBERTa [67], Electra [28], and DistilBERT [140].

For the baseline experiments, I employed a pre-trained BERT model as the base model, followed by one of two configurations:

1. An LSTM layer succeeded by a dense layer
2. A BiLSTM layer, followed by an LSTM layer and a dense layer

Throughout the training process, the weights of the BERT transformer remained frozen. The input text was fed into the transformer, from which we extracted the last hidden states, representing the embedded representation of each token. This output was then propagated to the subsequent layers. In models incorporating argumentative features, I augmented the transformer's last hidden states by concatenating them with a one-hot-encoded representation of argument components and relationships. This enriched feature representation was then input into the ensuing RNN-based layers. All models in our experiments utilized 0.2 as dropout and the Adam optimizer with PyTorch's default hyperparameters.

For the transformer-based models, I specifically chose pre-trained language models optimized for Token Classification tasks. These models are particularly well-suited for this research objectives, as they are designed to assign labels to individual tokens within a sequence, which aligns closely with the task of identifying fallacies in political debates, specifically. The primary distinctions among the selected models lie not only in their architectural differences, but also in the diverse corpora used for their pre-training. This variety in pre-training data allows to leverage models with potentially different learned representations and biases, which may impact their performance on this specific task.

The transformer models employed in this specific configuration are:

- **BertForTokenClassification**: Bidirectional approach for contextual information. Checkpoint employed:
  - `bert-base-uncased`
  - `bert-large-cased-finetuned-conll03-engl`
- **DebertaForTokenClassification**: Modified architecture with "de-coupled attention" and "cross-layer parameter sharing". Checkpoint employed:

  - `microsoft/deberta-base`

- **ElectraForTokenClassification**: Uses "discriminative pre-training" with generator and discriminator. Checkpoint employed:
  - `bhadresh-savani/electra-base-discriminator-finetuned-conll03-english`

- **DistilbertForTokenClassification**: Distilled version of BERT with reduced model size and computational needs. Checkpoint employed:
  - `distilbert-base-cased`
  - `distilbert-base-uncased`

In the initial experiments, the BertForTokenClassification (BERT FTC) model, utilizing the `bert-large-cased-finetuned-conll03-engl` checkpoint, showed the highest performance. Based on the results achieved by this model, I implemented MultiFusionBERT, an enhanced architecture designed to improve fallacious text detection. MultiFusionBERT integrates additional features including argumentative components labels(i.e., Claim and Premise), argumentative relations labels(i.e., Support and Attack), and Part-of-Speech (PoS) tags, as partially already done in Section 6.1.1. The inclusion of argumentative features was driven by the need to improve the model's understanding of argument structure, enabling it to better detect compromised logical structures — a typical feature of fallacious reasoning. I utilized the ElecDeb60to20 dataset's specific annotations for each fallacy, including argument component and relations labels. The use of PoS information was motivated by the observation that certain fallacies exhibit distinctive linguistic patterns. For instance, Loaded Language fallacies often employ emotionally charged adjectives and adverbs, while Ad Hominem fallacies typically use nouns or pronouns followed by negatively connoted adjectives. By leveraging PoS tagging, MultiFusionBERT is better equipped to identify these subtle linguistic cues associated with various fallacy types.

The MultiFusionBERT architecture, as illustrated in Figure 6.3, is designed for the detection and classification of fallacies in political debates. At its core, the model employs specialized TokenForClassification Transformer models to compute logits (L) for each feature type. These models are adapted to accommodate the specific number of labels for each feature: 3 for argumentative components and relations, and 17 for Part-of-Speech tags. To optimize parameter utilization, the architectures for argumentative features (components and relations) share the same parameters, allowing for the simultaneous computation of logits for both. A separate model, configured for the 17 PoS tags, generates logits for the PoS features. Figure 6.2 illustrates the encoding method applied to an example, demonstrating how the input is prepared for token label prediction by the architecture. Notably, the alignment of argument features with the tokenizer's offset mapping during tokenization is highlighted.

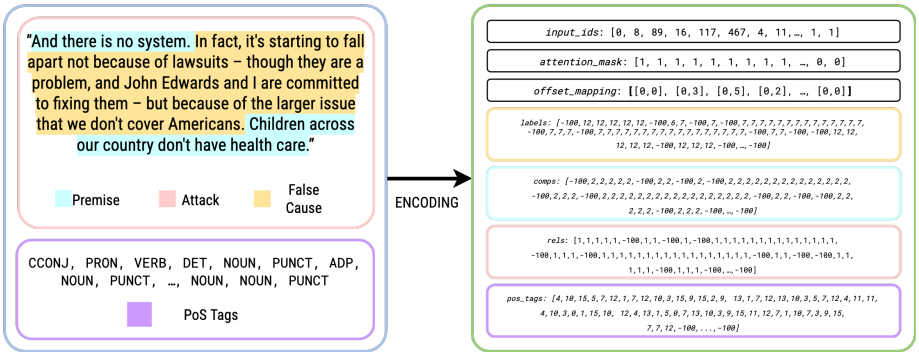


Figure 6.2: Encoded example of a single item of the dataset ElecDeb60to20. Figure drawn from [51].

Each argumentative feature is represented by a tensor of length 256 (the maximum token length of 216 is used) filled with label IDs (i.e., 0: *None*, 1: *Claim/Support*, 2: *Premise/Attack*) up to the maximum length. The same process is applied to the Part-of-Speech tensor, where each tag is converted into its corresponding ID (0: *ADJ*, 1: *ADP*, 2: *ADV*, 3: *AUX*, 4: *CCONJ*, 5: *DET*, 6: *INTJ*, 7: *NOUN*, 8: *NUM*, 9: *PART*, 10: *PRON*, 11: *PROPN*, 12: *PUNCT*, 13: *SCONJ*, 14: *SYM*, 15: *VERB*, 16: *X*).

Thus, the approach employed results in the computation of distinct losses for each aspect of the model: fallacy loss ($loss_{fal}$), component loss ($loss_{comp}$), relation loss ($loss_{rel}$), and part-of-speech loss ($loss_{pos}$). The model combines these individual losses into a unified average loss, termed the $joint_{loss}$, using the following formula (inspired by [160]):

$$joint_{loss} = \alpha * \frac{(loss_{fal} + loss_{comp} + loss_{rel} + loss_{PoS})}{N_{loss}} \tag{6.2}$$

N_{loss} represents the number of losses considered by the model, and α is an arbitrary weighting factor initially set to 0.1. To optimize performance, I conducted an empirical investigation to determine the most effective α value, the results of which are detailed in the next section, as shown in Table 6.8.

This $joint_{loss}$ approach allows MultiFusionBERT to simultaneously learn from multiple feature types, potentially leading to more robust and accurate fallacy detection and classification. The back-propagation function incorporates all losses, ensuring that the model's parameters are updated based on the comprehensive performance across all feature types. To optimize performance, I conducted ablation studies, systematically evaluating various feature combinations, as shown in Table 6.6. This process allowed to fine-tune the model and gain insights into the relative importance of different linguistic and argumentative features for fallacy detection and classification in argumentative text.

To evaluate MultiFusionBERT, I performed the experiments using the last version

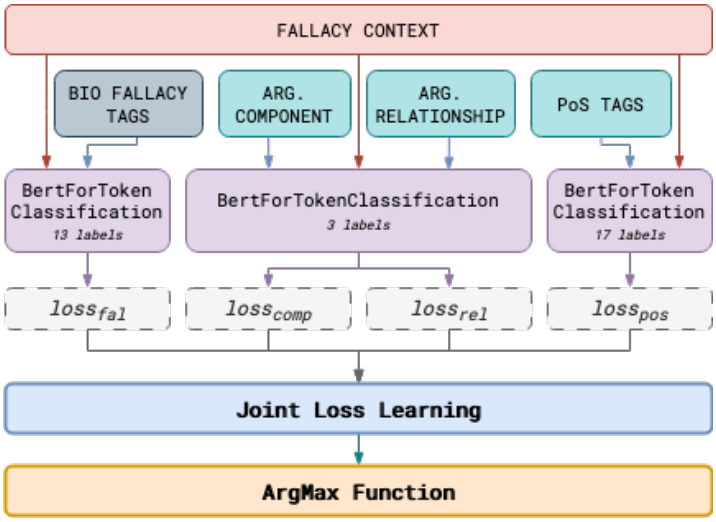


Figure 6.3: Architecture employed for the fallacy detection and classification task. Figure drawn from [50].

of the dataset, ElecDeb60to20, leveraging HuggingFace Transformers (v4.30) and PyTorch (v1.7.0). Models were fine-tuned using the Adam optimizer with a gradient clipping of 10, dropout of 0.1, and a learning rate of 4e-05. I used training and test batch sizes of 8 and 4, respectively, over 4 epochs. The train and test set split was performed considering the entire new dataset ElecDeb60to20. The new version of the dataset, ElecDeb60to20, was split 90-10 for training and testing. Fallacies are grouped in the following way: Appeal to Emotion (59.94%), False Cause (46.93%), Appeal to Authority (15.20%), Ad Hominem (13.58%), Slippery Slope (3.97%), and Slogans (2.63%). The final debate prominently featured an Appeal to Emotion and Ad Hominem fallacies, consistent with earlier debates. This trend was largely driven by discussions on the COVID-19 pandemic and candidates' personal issues, which, despite their disparate nature, dominated the debates. I employed scikit-learn's train_test_split function [127] with a random seed of 42 for reproducibility. Part-of-Speech tags, generated using spaCy, were incorporated to enhance linguistic understanding. The maximum tensor size was set to 256 tokens, ensuring comprehensive text coverage without truncation. This setup enabled a thorough assessment of MultiFusionBERT's efficacy in fallacy detection and classification within political debates.

6.2.2 Result and Discussion

Table 6.5 presents the comparative performance of all above-mentioned models in fallacy detection and classification within political debates. The evaluation metric employed is the macro-average F1 Score, which provides a balanced measure of precision and recall across the diverse fallacy categories: Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, Slogans, and Other (with B and I labels

merged for consistency). Despite the inherent challenges posed by the task's complexity and the relatively small dataset size, the results demonstrate promising advancements in automated fallacy detection and classification. Among the baseline models, BERT FTC (based on `conll` checkpoint) exhibited the highest performance, achieving an F1 Score of 0.7237. This result underlines the efficacy of pre-trained language models fine-tuned on domain-specific tasks. Notably, our proposed MultiFusion BERT architecture, which incorporates argumentative features (component and relation labels) and Part-of-Speech (PoS) tags, significantly outperformed all other models. With an F1 Score of 0.7394, MultiFusion BERT demonstrated a 2.17% performance increase over the best-performing baseline (BERT FTC `conll` checkpoint). This improvement, while modest in absolute terms, is substantial given the challenging nature of fallacy detection in nuanced political discourse. The great performance of MultiFusion BERT underscores the synergistic effect of combining deep contextual representations with task-specific features (argumentative structure and PoS information).

Model	**F1 Score**
BERT + LSTM	0.4697
BERT + LSTM (comp. and rel. features)	0.5142
BERT + BiLSTM + LSTM	0.5495
BERT + BiLSTM + LSTM (comp. and rel. features)	0.5614
BERT FTC (uncased)	0.7096
BERT FTC (conll)	0.7237
DeBERTaFTC deberta-base	0.7222
ElectraFTC (conll)	0.4033
DistilbertFTC (cased)	0.7010
DistilbertFTC (uncased)	0.7047
MultiFusion BERT (comp., rel. and PoS features)	**0.7394**

Table 6.5: Average macro F1 Scores for fallacy detection (BIO labels are merged) using different models. The scores are based on an average of 3 runs, except for BERT + (Bi)LSTM(s) models, which were evaluated using 10 runs. (FTC stands for "ForTokenClassification)

To assess the impact of different features on MultiFusion BERT's performance in fallacy detection, I conducted ablation tests using various combinations of argumentative components, relations, and Part-of-Speech (PoS) tags. Table 6.6 presents the average macro F1 Scores for these combinations. Individual features showed varying effectiveness, with PoS tags alone (F1 Score 0.7212) outperforming individual argumentative features (both at 0.6922). Among pairwise combinations, argumentative components and relations together yielded the highest score (0.7278). Notably, the integration of all three features resulted in the best performance (F1 Score 0.7394), significantly surpassing all other combinations and the baseline BERT FTC model (0.7237). These results highlight several key insights: *1)* The complementary nature of the features, as their full integration demonstrates a synergistic effect beyond individual contributions, *2)* The importance of linguistic structure, evidenced by the strong performance

of PoS tags, and *3)* The value of capturing complete argumentative context, shown by the superior performance of component-relation pairs.

Features			Avg macro
Comp.	**Rel.**	**PoS**	**F1 Score**
✓			0.6922
	✓		0.6922
		✓	0.7212
✓	✓		0.7278
✓		✓	0.7166
	✓	✓	0.7166
✓	✓	✓	**0.7394**

Table 6.6: Average macro F1 Scores for fallacy detection (BIO labels are merged) using MultiFusion BERT and different features. The scores are based on an average of 3 runs.

Table 6.7 presents the detailed classification report for MultiFusion BERT, revealing varied performance across fallacy types in political debates. The model excels in identifying 'I-' (Inside) labels for Ad Hominem (F1 0.88), Appeal to Authority (F1 0.84), and False Cause (F1 0.84), suggesting effective capture of fallacious argument continuations. It also performs strongly in distinguishing non-fallacious segments ('O' category, F1 0.93). However, challenges persist in detecting Slogans (F1 0.00) and identifying the starting point of fallacious arguments ('B-' labels generally). The model generally demonstrates higher precision than recall, indicating a conservative prediction approach. The macro average F1 Score of 0.56 reflects the challenge of balancing performance across all fallacy types, while the weighted average F1 Score of 0.88 shows strong overall performance when accounting for label frequencies.

The joint loss function of MultiFusion BERT, defined in Equation 6.2, was explored with α values of 0.1, 0.3, and 0.5 across various feature combinations. Table 6.8 reveals that α's impact varies significantly depending on the features used, with the highest macro F1 Score (0.7394) achieved at $\alpha = 0.1$ when all features are integrated. Notably, PoS consistently outperforms other single features, while the Comp. + Rel. combination generally excels among two-feature sets, particularly at lower α values. The inconsistent performance improvements across different α values and feature combinations highlight the complexity of tuning multitask learning models for fallacy detection and classification. These findings underscore the importance of careful hyperparameter optimization, suggesting that the optimal α is dependent on the specific feature set employed.

6.2.3 Error Analysis

Table 6.9 presents a comprehensive analysis of MultiFusion BERT's performance on the test set, considering the merged labels in the fallacy detection and classification task. The confusion matrix in Figure 6.4 provides additional insights into the model's prediction patterns. In examining performance across labels, it's notable that the model

Label	Prec.	Rec.	F1 Score	Support
B-AdHominem	1.00	0.19	0.31	27
B-AppealtoAuthority	0.75	0.50	0.60	30
B-AppealtoEmotion	0.72	0.39	0.51	120
B-FalseCause	0.75	0.33	0.46	9
B-Slipperyslope	0.33	0.12	0.18	8
B-Slogans	0.00	0.00	0.00	5
I-AdHominem	0.98	0.79	0.88	712
I-AppealtoAuthority	0.90	0.78	0.84	1,019
I-AppealtoEmotion	0.81	0.78	0.79	2,104
I-FalseCause	0.81	0.87	0.84	312
I-Slipperyslope	0.89	0.89	0.89	324
I-Slogans	0.00	0.00	0.00	44
O	0.90	0.95	0.93	7,914
Accuracy			0.88	12,628
Macro avg	0.68	0.51	**0.56**	12,628
Weighted avg	0.88	0.88	0.88	12,628

Table 6.7: Classification report of MultiFusion BERT considering the single labels predicted for each token.

$\alpha = 0.1$				$\alpha = 0.3$				$\alpha = 0.5$			
Comp.	**Rel.**	**PoS**	**F1**	**Comp.**	**Rel.**	**PoS**	**F1**	**Comp.**	**Rel.**	**PoS**	**F1**
✓			0.6922	✓			0.7054	✓			0.7057
	✓		0.6922		✓		0.7054		✓		0.7057
		✓	0.7212			✓	**0.7214**			✓	0.6817
✓	✓		0.7278	✓	✓		0.6889	✓	✓		**0.7366**
✓		✓	0.7166	✓		✓	0.7160	✓		✓	0.7054
	✓	✓	0.7166		✓	✓	0.7160		✓	✓	0.7054
✓	✓	✓	**0.7394**	✓	✓	✓	0.7084	✓	✓	✓	0.7070

Table 6.8: MultiFusion BERT's average macro F1 Scores for fallacy detection and classification, comparing feature combinations (Comp., Rel., PoS) and α values (0.1, 0.3, 0.5). Scores averaged over 3 runs; B and I labels are merged.

exhibits the poorest performance in identifying tokens labeled as Slogans, despite this category being relatively easy for human recognition. This lack of performance can be attributed to limited examples in both training and test sets, as well as the complexity of recognizing slogans in political debates, which often involves semantic and pragmatic factors beyond the syntactic and argumentative features emphasized by the model. The model consistently struggles with accurately identifying both "B-Slogans" and "I-Slogans" tags (see Table 6.7). On the other hand, the categories of Slippery Slope and False Cause present the highest performance metrics, with F1 Scores of 0.89 and 0.84, respectively. This success may be due to the well-defined structure of Slippery Slope arguments, which often portray improbable or exaggerated consequences, and the model's ability to capture both argumentative components and semantic nuances associated with these fallacy types. Performance on remaining labels aligns with

Merged Label	Prec.	Rec.	F1
Ad Hominem	0.99	0.77	0.87
Appeal to Authority	0.90	0.78	0.83
Appeal to Emotion	0.82	0.77	0.79
False Cause	0.82	0.86	0.84
Slippery Slope	0.90	0.88	0.89
Slogans	0.00	0.00	0.00
O	0.90	0.95	0.93
Accuracy			**0.89**
Macro avg	0.76	0.72	0.74
Weighted avg	0.89	0.89	0.89

Table 6.9: MultiFusion BERT classification report of fallacy detection and classification task. The *B* and *I* labels are merged.

previous findings by [51] for the classification task, suggesting consistency in fallacy detection across different debate contexts.

Furthermore, the analysis of the normalized confusion matrix (Figure 6.4) reveals several important patterns. Despite the "O" class (non-fallacies) having the highest F1 Score, the model tends to over-predict instances in this category, as evidenced by the higher proportion of false positives in the "O" column. In terms of misclassification patterns, False Cause and Appeal to Emotion are most frequently misclassified as non-fallacious, while a notable proportion of Appeal to Authority instances are misclassified as Appeal to Emotion. The use of a normalized confusion matrix helps mitigate the impact of class imbalance in the dataset, providing a clearer picture of the model's relative performance across categories. To further analyze the model's performance, I examined specific instances of misclassification.[3]

Example 6.2.1. **Franklin Roosevelt said in 1932 that the only thing we have to fear is fear itself.** Appeal to Authority Appeal to Emotion

Example 6.2.1 demonstrates a case where the model misclassifies an Appeal to Authority as an Appeal to Emotion. This error likely comes from the presence of the emotionally charged word "fear", which appears to interfere with the authoritative nature of the quote's attribution to Franklin Roosevelt. This suggests that the model may be overly sensitive to emotional lexicons, potentially at the expense of recognizing appeals to credible sources.

Example 6.2.2. **As the President said the other night, there will always be troubles in this ol' world, but the United States of America** *can be counted on to provide the vision that the world looks for from the United States of America.* Other Appeal to Authority + Other

In Example 6.2.2, we observe an instance of false positive identification. The model incorrectly labels a non-fallacious statement (labeled as 'Other' in the ground truth) as

[3] Underlined text is to highlight the true label for each token, whereas **Bold** is for the predicted fallacy and *Italic* for predicted *O* tokens.

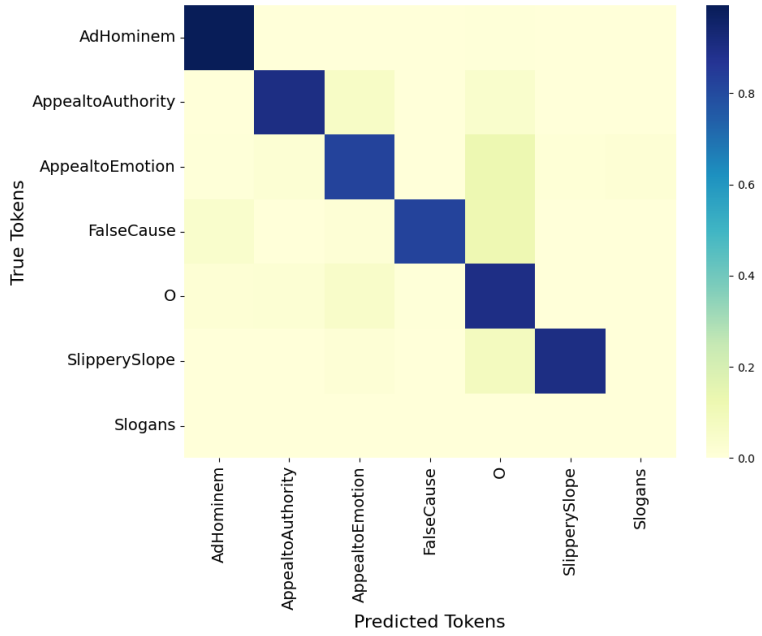


Figure 6.4: Normalized confusion matrix of MultiFusion BERT. BIO labels are merged. Normalization is performed using the number of true elements in each class.

an Appeal to Authority. This misclassification indicates that the model may be overgeneralizing the concept of authority appeals, possibly due to the mention of "the President" in the text.

Example 6.2.3. **But as Admiral Yarnell has said, and he's been supported by most military authority, these islands that we're now talking about are not worth the bones of a single American soldier;** *and I know how difficult it is to sustain troops close to the shore under artillery bombardment.* Appeal to Authority Appeal to Emotion + Other

A significant misclassification occurs in Example 6.2.3, where the true label for the entire statement is Appeal to Authority, but the model incorrectly predicts Appeal to Emotion for part of the text and Other (non-fallacious) for the remainder. This error highlights the model's difficulty in recognizing Appeals to Authority, especially involving military or technical expertise, and suggests an oversensitivity to emotional language. It also reveals inconsistency in classification across a single statement, implying that the model may rely too heavily on specific keywords rather than grasping the overall rhetorical structure.

Example 6.2.4. **In a place like Chicago, where thousands of people have been killed,** *thousands over the last number of years, in fact, almost 4,000 have been killed since Barack Obama became president, overall almost 4,000 people in Chicago have been killed. We have to bring back law and order.* False Cause False Cause + Other

Lastly, Example 6.2.4 illustrates the model's partial success in identifying a False

Cause fallacy. While it correctly detects the fallacious reasoning in part of the statement, it fails to extend this classification to the entire relevant portion. This partial recognition suggests that the model may have difficulty in determining the full scope of fallacious arguments, especially when they are embedded within longer statements. These examples collectively illustrate the model's strengths and limitations in fallacy detection.

6.3 Summary

This chapter presents a systematic study, firstly, on fallacy classification, and then on fallacy detection and classification in political debates. The researches are motivated by the limitations of traditional argumentation schemes when applied to real-world scenarios, particularly in political debates, where fallacious arguments often closely resemble sound reasoning. The chapter details several key contributions and findings. The fallacy classification task was significantly enhanced using the ElecDeb60to16 dataset, enriched with fallacy annotation labels. By employing the Longformer model, which excels at processing long text sequences, and integrating argumentative labels, I achieved a substantial improvement in multiclass classification performance. The F1 Score increased from 0.61 to 0.84, demonstrating the effectiveness of our approach in identifying fallacious arguments in political debates and integrating elements of context. These results from the fallacy classification task lay a strong foundation for the subsequent fallacy detection and classification task. The significant improvement in F1 Score indicates that the approach not only enhances the identification of fallacies but also provides a more nuanced understanding of their context within political debates. Thus, the ElecDeb60to16 dataset has been expanded by incorporating the 2020 Trump-Biden presidential debate, complete with argumentative annotations and fallacy labels. This extension enhances the corpus's representativeness and allows for more robust analysis. A major innovation introduced in this chapter is the MultiFusion BERT, a novel transformer-based architecture that integrates debate text, argumentative features (components and relations), and engineered features. This innovative approach significantly improves both fallacy detection and classification tasks. The MultiFusion BERT model achieved an average performance increase of 2.12% compared to baseline methods and competing approaches, underscoring the effectiveness of incorporating argumentative features in fallacy analysis. The success of the MultiFusion BERT model in the fallacy detection and classification task builds upon the insights gained from the earlier fallacy classification task. By integrating argumentative features and engineered features, this approach not only improves performance but also demonstrates the value of combining different aspects of argumentation analysis. This progression from classification to detection and classification represents a significant advancement in the field of computational argumentation, particularly in the context of political debates.

Chapter 7

Repairing Fallacies in Political Debates

This chapter explores an innovative application of Argument Mining in political debates, focusing on argumentative fallacies. The dataset employed to evaluate the ability of the LLMs is named FallacyFix: A Repaired Fallacies Dataset, the first dataset made of human-repaired fallacious arguments. Thus, this chapter investigates the capability of various Large Language Models (LLMs) to repair and classify fallacies in political debates, including categories such as Appeal to Fear, Appeal to Pity, Appeal to Popular Opinion, Flag Waving, and Loaded Language. The experiments are driven by a combination of settings and configurations based on a modular prompt given to the LLMs, allowing for a systematic exploration of model performance under different conditions. Thus, this research evaluates LLM performance using both automated metrics like BERTScore and human evaluation, comparing model-generated repairs against gold standards.

This chapter explores an innovative application, focusing on political debates and, in particular, on argumentative fallacies previously discussed in Section 2.4. Scientific literature has responded to the challenge of counter the spread of fallacious and propagandistic arguments by proposing various methodologies for identifying such arguments in texts. Recent works by [138, 23, 99, 69, 126, 100, 51, 50] have made significant progress in this area, employing techniques ranging from traditional machine learning to advanced natural language processing models. These studies have shown promising results in automating the detection of fallacies across different contexts and languages. However, the mere classification (Section 6.1) or detection and classification (Section 6.2) of argumentative fallacies in political debates, while valuable, is insufficient to address the crucial problem of fallacies. These approaches do not ensure that the public fully understands the impact of such arguments on their decision-making process, nor do they adequately develop critical thinking skills. In simple terms, identifying a fallacy is only the first step; the real challenge lies in understanding its impact and learning how to construct more valid arguments. Therefore, this chapter addresses

a more ambitious challenge [94, 16, 159]: not only identifying fallacies but also explaining why a particular argument is considered fallacious and demonstrating how it can be **repaired** into a valid, non-fallacious argumentation. This approach is inspired by research in cognitive psychology and education, which suggests that active engagement with flawed reasoning can significantly enhance critical thinking skills [1, 125]. To address this critical challenge, I introduce a novel task named *fallacious argumentation repair*. This task aims to transform statements containing fallacious arguments into versions that are *clearer*, *fairer*, and *free from any techniques that could negatively persuade listeners*. The concept of *repair* here is not merely about correcting logical structure, but also about enhancing the overall quality and ethical standing of the argument. This task has been implemented within the context of political debates, where the necessity for such a solution is particularly pressing. Political debate shapes public opinion, influences political decisions, and ultimately affects the functioning of democratic societies. By providing a method to repair fallacious arguments, this research aims to elevate the quality of political debate and, by extension, the democratic process itself. By doing so, this research not only contributes to the academic discourse on argument analysis but also provides a practical tool for enhancing the quality of public political debates. It bridges the gap between theoretical understanding of fallacies and practical application in real-world contexts. Moreover, it opens up new avenues for research in areas such as automated fact-checking, political communication studies, and the development of educational tools for critical thinking.

The structure of this chapter is as follows: Section 7.1 outlines the research scope and presents the corresponding research questions. Section 7.2 describes the methodologies and models used for fallacy repair. Section 7.3 presents the experimental results, while Section 7.4 provides a comprehensive error analysis to enhance the understanding of these findings.

7.1 Fallacy Repair

This research aims to address the challenge of repairing fallacious arguments in political debates, casting it as a computational task and assessing the performance of large language models in comparison to human annotations in generating repaired arguments. The process of repairing fallacies involves transforming statements containing logical flaws into versions that are more precise, balanced, and free from persuasive techniques that may mislead audiences. While this task can be highly subjective in principle, in Section 7.1.1 I proposed a methodology to ensure that the repaired statements are impartial and devoid of manipulative rhetoric that is present in fallacies [40, 43]. It is important to note that currently, there exist no standardized guidelines for transitioning an argument from fallacious to non-fallacious, nor is there a universally accepted definition of fallacy repair. This lack of formal criteria underscores the complexity and novelty of the task at hand.

To enhance the primary objective of fallacy repair, I included a sub-task that prompts language models to classify also the fallacy type before repairing it. By comparing the models' performance when the fallacy label is given as input versus when it is omitted, we can evaluate both their ability in fallacy classification and in the repairing task, in order to provide a more comprehensive assessment of the models' capabilities in handling fallacious arguments under different scenarios.

Thus, this research addresses several interrelated research questions to explore the potential of large language models in the automatic repair of fallacious arguments. In scenarios where explicit labels are not provided, I investigate whether LLMs are capable of accurately identifying fallacy types (**RQ1**). Furthermore, I examine the efficacy of LLMs in repairing fallacious arguments within political debates (**RQ2**), comparing their performance to human annotations in transforming these arguments into clearer, fairer statements devoid of manipulative rhetoric [40, 43]. Lastly, I am interested in determining if the repaired arguments not only enhance human comprehension but also offer additional insights beyond the original input, thereby evaluating the potential of LLMs to contribute meaningfully to the improvement of argumentative discourse (**RQ3**).

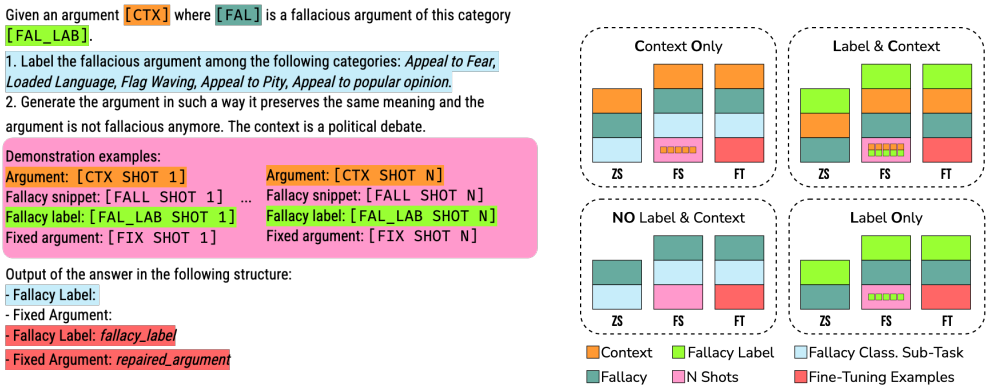
To thoroughly address these research questions, I created a new dataset named FallacyFix that includes a set of human-repaired fallacious arguments, as detailed in Section 3.2.4. Thus, I implemented a modular approach, outlined in Section 7.1.1, which incorporates various input components to drive the model's prediction and repair processes.

7.1.1 Prompt's Configuration

This research employs large language models to autonomously generate human-like arguments that address and repair the fallacious nature of arguments in political debates. The generation process is guided by a meticulously crafted instruction prompt, designed to yield LLM outputs that can be directly compared to the gold standard expert arguments annotated, described in Section 3.2.4. The experimental design carefully involves two crucial components: *(1)* the fallacy label (`L`) and *(2)* the contextual information (`C`) surrounding the fallacy to be repaired, allowing for a comprehensive analysis of their impact on the repair process.

The experimental design employed to build the prompt consists of Zero-Shot, Few-Shot, and Fine-Tuning scenarios, each applied across four distinct configurations: `LC` (Label and Context), `CO` (Context Only), `LO` (Label Only), and `NO` (Neither Label nor Context). For the `LO` and `NO` configurations, I employed a subset of the FallacyFix dataset. This subset comprised 541 examples where fallacies were rectified through paraphrasing or partial text modification, a reduction from the original 747 repaired examples. I excluded instances where the repair method involved the complete removal of the original fallacious text. This selection criterion was necessary because classifying

a fallacious snippet in the prompt becomes unfeasible when the repaired version has entirely eliminated the original text.



(a) Textual representation. **(b)** Minimalistic representation.

Figure 7.1: Different representations of the instructed prompt given to the LLMs to repair the fallacy.

The prompt configuration, as illustrated in Figure 7.1, adopts a modular approach with color-coded components serving distinct functions. When present, the **orange** section provides contextual information [CTX], while the **pine green** section contains the core fallacious argument [FAL], which forms the basis of the analysis. The handling of fallacy labels is managed through mutually exclusive components: the **green** section is utilized when a label is provided [FAL_LAB], whereas the **cyan** section requests a label for the fallacy classification sub-task when one is not given. The complexity of the prompt scales across different experimental configurations:

- **Zero-Shot**: Utilizes only the base components of the prompt.
- **Few-Shot**: Incorporates demonstrative examples (denoted in **magenta**) such as [CTX SHOT 1], [FALL SHOT 1], [FALL_LAB 1], and [FIX SHOT 1].
- **Fine-Tuning**: Integrates additional training instructions (highlighted in **red**) to further refine the model's performance.

This modular approach provides a comprehensive evaluation of LLM performance across various experimental conditions, yielding insights into the models' capacities for fallacy classification and repair under different levels of contextual and label information. By systematically manipulating these input variables, I can quantify the impact of each component on the quality and accuracy of the generated repairs. The experimental design provides a detailed analysis of the correlation between diverse input parameters and the resulting model outputs, clarifying the extent to which LLMs can emulate sophisticated human-like reasoning in the context of fallacious argument analysis and improvement.

7.2 Experimental Setup

To address the aforementioned tasks, I employed a diverse set of widely used large language models . Additionally, I included a baseline model for comparative analysis. This baseline served to benchmark the performance of LLMs in predicting fallacy labels and generating repaired arguments from fallacious ones. The evaluated LLMs in this study are as follows:

- **BART** [98]: Based on the `facebook/bart-large` pre-trained checkpoint.
- **Google Gemma** [150]: Utilizing the `google/gemma-1.1-7b-it` and `google/gemma-1.1-2b-it` model variants.
- **Mistral 7B** [78]: Employing the `mistralai/Mistral-7B-Instruct-v0.2` version.
- Mixtral 8x7B [79]: Using the `mistralai/Mixtral-8x7B-Instruct-v0.1` configuration.
- **LLaMA 3** [3]: Tested with the `meta-llama/Meta-Llama-3-8B` model release.

All the aforementioned models are open-source and were accessed via the Hugging Face platform[1]. In addition to these, I evaluated two proprietary language models:

- **OpenAI GPT**: Utilizing the `gpt-3.5-turbo-0125` and `gpt-4-0125` versions, accessed through the OpenAI API[2].
- **Claude** [7]: Employing the high-performance `claude-3-opus-20240229` version, accessed via the Anthropic API[3].

7.2.1 Metrics

To assess the quality of the generated repaired arguments, this research employs a combination of quantitative and qualitative evaluation methods.

Automatic Evaluation Metrics

The quantitative assessment uses standard text completion metrics, which evaluate both lexical and semantic similarities between the LLM-generated text and the gold-standard repaired text provided by human annotators. This approach accounts for various methods of repairing fallacious arguments, such as paraphrasing, omission, generalization, or weakening, while ensuring that the core intent and clarity of the original argument are maintained. The following automatic metrics are employed in our evaluation:

- **BERTScore** [173]: This metric evaluates the similarity between generated and reference texts using contextual embeddings from the BERT model to capture semantic meaning[4]. BERTScore has been chosen for its ability to capture semantic

[1] `https://huggingface.co/`
[2] `https://openai.com/api/`
[3] `https://www.anthropic.com/api`
[4] `https://huggingface.co/docs/evaluate/`

nuances in context, making it particularly suitable for evaluating fallacy repairs where subtle wording changes are crucial. This contextual sensitivity addresses the lack of standardized methods for comparing repaired fallacious arguments.

- **IOU-F1**: Combining the Intersection over Union (IoU) with the F1 score, this metric measures the overlap between predicted and true labels. IOU-F1 assesses both precision and recall, providing a balanced view of the model's performance in identifying and repairing fallacies.
- **Macro Average F1 Score** for Fallacy Classification: This metric offers a balanced assessment of model performance across all fallacy labels. It is especially crucial in this research due to potential class imbalances in the dataset, ensuring that the evaluation is not dominated by more frequent fallacy types.

These metrics were carefully selected for their robust methodologies, in alignment with the task objectives, and with the ability to provide meaningful insights into the results. They offer complementary perspectives on the quality of the generated repairs, capturing both fine-grained semantic similarities and broader classification accuracy.

Human Evaluation Metrics

While automatic metrics provide valuable quantitative insights, they may not fully capture the nuanced aspects of argument repair that require human judgment. To address this limitation, the study incorporates a set of human-based metrics, enriching the automatic evaluation with qualitative assessments. Drawing from prior research in natural language generation, argument quality assessment, and AI-generated text evaluation, a holistic human evaluation framework was developed. This framework synthesizes insights from several key studies to address the unique challenges of assessing fallacy repair in AI-generated text. The work of Sourati et al. (2023) [144] emphasized the importance of robust, multidimensional evaluation criteria that capture both the logical structure and linguistic quality of arguments, which informed my approach to holistic assessment. Following up on this foundation, elements from [29] on evaluating AI-generated explanations were incorporated, adapting their focus on relevance, coherence, and factual correctness to our fallacy repair context. [170] provided valuable insights into assessing cogency and soundness, which are crucial aspects of effective fallacy repair. The resulting human evaluation framework assesses the repaired fallacious arguments using a 5-point Likert scale for each of the following criteria:

- **Relevance**: This metric evaluates whether the repaired fallacious argument aligns with the topic and category of the original fallacy. A high relevance score indicates that the repair maintains the core subject and argumentative context of the original fallacy.
- **Suitability**: This criterion assesses the appropriateness of the repaired fallacy's style, including aspects such as politeness, neutrality, and avoidance of explicit references while maintaining the original meaning. A suitable repair should address the fallacious reasoning without introducing new issues of tone or content.

- **Cogency**: This metric measures the logical correctness and coherence of the repaired fallacy with respect to the given prompt. A cogent repair should present a logically sound argument that effectively addresses the fallacious reasoning in the original text.

To optimize the evaluation process, a pilot study was conducted to assess time requirements and establish guidelines[5]. Based on the findings, an approach favoring a larger annotator pool over multiple examples per annotator was adopted. This decision was driven by the task's complexity, with each annotator requiring approximately 90 minutes to evaluate 15 fallacies. This strategy aimed to maximize dataset diversity within resource constraints, prioritizing a broader range of perspectives over repeated evaluations. The choice to prioritize a larger annotator pool over multiple annotations per example was primarily motivated by the novelty of the task. As a proof of concept, it was considered crucial to gather insights from a diverse group of annotators, given the unprecedented nature of evaluating LLM-repaired fallacious arguments. This approach allows for the capture of a wider range of human judgments and potentially identifies patterns or discrepancies in how different individuals perceive the quality of repaired arguments. A total of 17 volunteer annotators were recruited, who participated after providing informed consent through a standardized form. To ensure consistency and clarity in the evaluation process, detailed instructions were provided to the annotators. These instructions were designed to give context to the task and provide a clear structure for the evaluation process. The key components of the instructions were as follows:

- **Context**: Annotators were informed that the project focuses on the generation of repaired political arguments using large language models . They were told that while automated metrics such as BERTScore and IOU-F1 had been used to evaluate the generated text, the goal of the human evaluation was to determine if the generated arguments align with those created by humans in terms of Relevance, Suitability, and Cogency of fallacy repair.
- **Task Instructions**: The annotators were given a step-by-step guide to complete their evaluation:
    1. Complete an anonymous demographic survey to provide context for the analysis.
    2. Access the provided Google Sheets document containing the evaluation materials.
    3. Navigate the document, which was structured with seven columns:
        - Column 1: The original prompt or fallacious argument
        - Column 2: The raw generated answer from the LLM
        - Column 3: The extracted repaired argument

[5]The annotation guidelines used in the process are provided here.

– Columns 4-7: The evaluation criteria to be filled out by the annotator

The experimental design, as outlined in Section 7.1.1, covered various configurations to thoroughly evaluate the performance of large language models in repairing fallacious arguments. To overcome GPU memory constraints while maintaining the ability to work with these sophisticated models, I employed the Low-Rank Adaptation (LoRA) technique [72]. This approach enabled efficient loading and fine-tuning of LLMs without exceeding hardware limitations. Thus, to ensure comparability across all experiments, I maintained consistent hyperparameters for all LLMs, regardless of the specific configuration. In both Zero-Shot and Few-Shot settings, I set the temperature to 0.5, striking a balance between creative output and logical coherence. The maximum token count was fixed at 512 to encourage concise responses while allowing sufficient space for comprehensive argument repair. To optimize computational efficiency while thoroughly exploring LLM capabilities, I limited the number of generated responses to 1 per prompt. The decision to generate a single response per prompt was based on methodological and computational considerations. Implementing a multi-response approach would have necessitated an additional selection algorithm, significantly increasing computational complexity and potentially introducing confusing variables. Given the study's extensive variety of models, configurations, and prompts, this approach was considered optimal for maintaining experimental consistency while ensuring reliable model evaluation across various conditions. I left the stop parameter unset (None), allowing each model to autonomously determine the natural endpoint of its responses. Furthermore, for the Few-Shot setting, I implemented a carefully designed sampling strategy. I randomly selected one example from each of the five fallacy subcategories involved in the experiment. This approach to demonstration examples was motivated by pilot experiments, which indicated that this number effectively balances efficiency and output quality. My decision aligns with previous research in the related domain of hate speech detection, where studies have shown that providing 5–10 examples is generally sufficient for optimal LLM performance [124, 65]. The connection between hate speech and fallacious arguments is rooted in their shared use of biased or deceptive language to unfairly influence opinions or target specific groups [111]. The Fine-Tuning process involved additional parameters to further refine my experimental approach. I set the maximum sequence length to 1024 tokens, optimizing the processing of both input and output texts. The training consisted of 3 epochs, with a learning rate of 2e-4 and a weight decay of 0.01. I utilized the paged_adamw_32bit optimizer, a choice that balances efficiency and effectiveness in training large language models. To manage computational resources effectively, I set the training batch size to 4 and the evaluation batch size to 2.

7.3 Results & Discussion

To address **RQ1**, I evaluated the ability of various large language models to accurately categorize fallacies using Zero-Shot (ZS), Few-Shot (FS), and Fine-Tuned (FT)

settings. The dual-task approach demonstrates the versatility and efficiency of LLMs in simultaneously performing fallacy classification and repaired text generation. This methodology differs this work from previous studies that relied on standard classification models [51, 60, 81, 157], contributing significantly to the advancement of automatic detection and defusion of fallacious arguments in political debates. Table 7.1 presents a complete overview of the models' performance, displaying the macro average F1 Score for fallacy label prediction across two configurations: Context Only (`CO`) and No Fallacy Label & Context (`NO`). These configurations allow assessing the models' capabilities under different information constraints. This analysis reveals that GPT-4 demonstrates superior performance in both ZS and FS settings across both configurations. In the `CO` configuration, GPT-4 achieves significant F1 scores of 43.48% and 59.15% for ZS and FS, respectively. Similarly, in the `NO` configuration, it maintains its lead with 26.32% (ZS) and 37.69% (FS). This consistent performance underscores GPT-4's strong out-of-the-box reasoning abilities in fallacy classification tasks. Interestingly, LLaMA 3 8B exhibits exceptional performance in the FT setting, significantly outperforming other models with F1 scores of 52.76% (`CO`) and 31.06% (`NO`). These scores demonstrate LLaMA 3 8B's strong capacity for task-specific adaptation when fine-tuned on fallacy classification data. The significant contrast between LLaMA 3 8B's FT performance and its ZS/FS results highlights the potential for considerable improvements through targeted training. Claude 3 shows consistent performance across ZS and FS settings in the `CO` configuration (37.94% and 37.22%, respectively), but exhibits a notable drop in the `NO` configuration (18.55% ZS, 23.41% FS). This indicates that Claude 3 may be more sensitive to the presence of explicit fallacy labels and context, an important consideration for real-world applications where such explicit information might not always be available. The baseline BART model, evaluated only in the FT setting, achieves F1 scores of 34.92% (`CO`) and 42.20% (`NO`), providing a useful reference point for assessing the performance gains of more advanced models. Notably, in the `NO` configuration, BART achieves the highest score, ranking first among all models, even though the method employed to perform the multi-class classification is different from the other models. This performance is particularly impressive given BART's status as a baseline model and highlights the effectiveness of its fine-tuning approach in this specific context. Other models like Gemma 1.1 2B, Gemma 1.1 7B, Mistral 7B, and Mixtral 8x7B generally show lower performance, particularly in ZS and FS settings. However, their FT performance suggests potential for improvement with task-specific training, which could be valuable in resource-constrained environments or applications requiring smaller model footprints. These results underscore the varying capabilities of different LLMs in fallacy classification tasks. The superior performance of GPT-4 in ZS and FS settings highlights its strong inherent reasoning abilities, while LLaMA 3 8B's notable FT performance demonstrates the potential for significant improvements through task-specific fine-tuning. Additionally, the impressive performance of the BART model in the FT setting, particularly in the `NO` configuration where it achieves the top score, underscores the importance of considering diverse approaches and model architectures in tackling complex language understanding tasks like fallacy classification.

	Context Only (CO)			No Fall. label & Context (NO)		
Model	**ZS**	**FS**	**FT**	**ZS**	**FS**	**FT**
BART	-	-	*34,92%*	-	-	*42,20%*
Claude 3	37,94%	37,22%	-	18,55%	23,41%	-
Gemma 1.1 2B	6,23%	2,35%	3,70%	5,22%	1,44%	4,88%
Gemma 1.1 7B	2,36%	0,91%	15,58%	1,54%	0,84%	6,69%
GPT 3.5	12,74%	26,68%	22,95%	7,44%	30,52%	13,33%
GPT 4	43,48%	**59,15%**	-%	26,32%	**37,69%**	-%
LLaMA 3 8B	6,68%	2,80%	52,76%	6,64%	6,82%	31,06%
Mistral 7B	0,00%	0,44%	-%	0,54%	0,59%	-%
Mixtral 8x7B	0,79%	1,05%	12,96%	0,92%	1,34%	4,48%

Table 7.1: Performance comparison of LLMs in fallacy classification based on macro avg F1 Score across different prompting strategies and contexts: *Zero-Shot* (ZS), *Few-Shot* (FS), and *Fine-Tuning* (FT) in CO and NO configurations.

To address **RQ2**, I compared the arguments generated by the LLMs with the gold standard collected and described in Section 3.2.4. Given the current absence of standardized evaluation protocols in this domain, we utilized BERTScore for semantic-based automatic evaluation and IOU-F1 for token-based evaluation. The initial analysis revealed that human-repaired fallacies closely match the original arguments, with a BERTScore of 0.91 ± 0.06. Different human repairs showed even higher similarity (0.94 ± 0.06). Notably, human repairs outperformed those by LLMs when compared to the original fallacies, suggesting that AI systems still fall behind humans in this complex task. Figures 7.2, 7.3, and 7.4 present heatmaps of the full metrics ablation results for Zero-Shot, Few-Shot, and Fine-Tuning settings, respectively. These visualizations offer insights into model performance across different configurations and metrics.

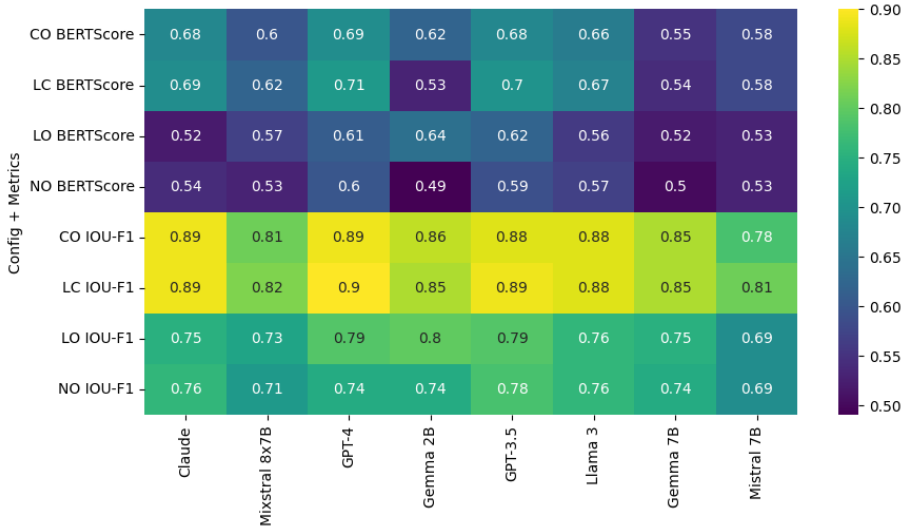


Figure 7.2: Heatmap of full metrics ablation study for the *Zero-Shot* setting.

In Figure 7.2, GPT-4 and Claude consistently outperform other models across most

configurations. IOU-F1 scores are generally higher than BERTScores, with `CO` and `LC` configurations showing the best performance. The `NO` configuration typically yields the lowest scores, highlighting the importance of context.

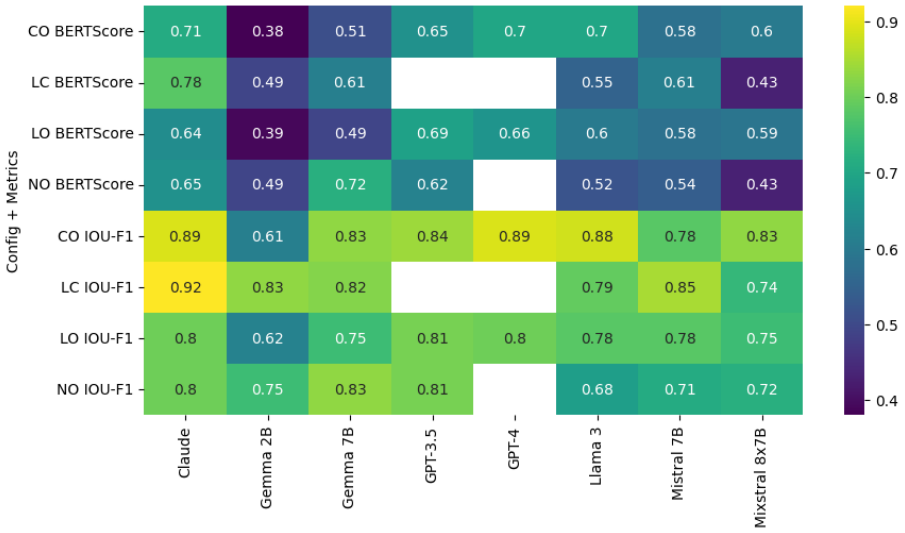


Figure 7.3: Heatmap of full metrics ablation study for the *Few-Shot* setting.

In Figure 7.3, Claude demonstrates superior performance, particularly in BERTScore metrics. GPT-4 maintains strong performance, especially in IOU-F1 scores. There's a noticeable improvement in scores compared to the Zero-Shot setting, indicating the benefit of Few-Shot learning.

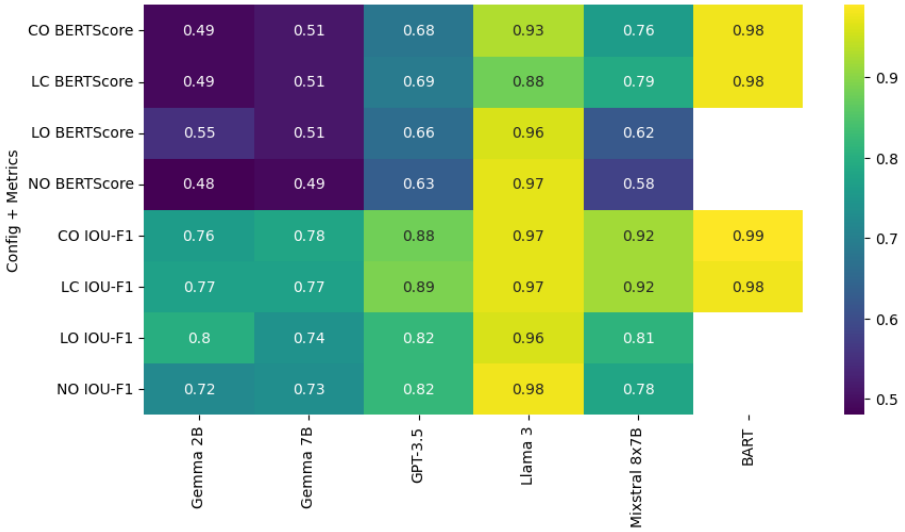


Figure 7.4: Heatmap of full metrics ablation study for the *Fine-Tuning* setting.

In Figure 7.4, LLaMA 3 shows remarkable improvement, consistently achieving the highest scores across all configurations, with average scores of 0.94 for BERTScore and 0.97 for IOU-F1. BART, included as a baseline, achieves the best results in both `CO` and

NO configurations, highlighting the effectiveness of fine-tuning for this task. It's worth noting that the method used with BART for fallacy label classification and general text generation differs significantly from other LLMs due to its unique prompting requirements. Despite these differences, BART's strong performance in the Fine-Tuning setting establishes a robust baseline for comparison with other models. The performance gap between different models narrows in this setting, suggesting that fine-tuning can significantly enhance capabilities across various architectures.

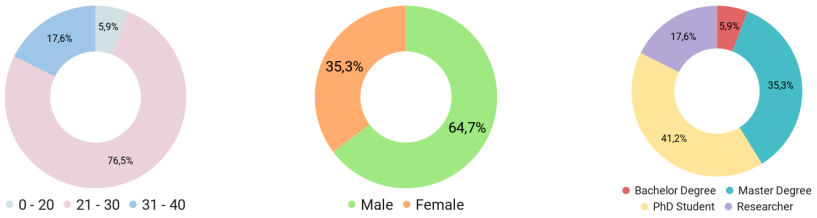
Across all settings, it is possible to observe that configurations without fallacy context (LO, NO) generally result in lower metric values, while those with context (CO, LC) exhibit higher values. This trend is particularly pronounced in Zero-Shot and Few-Shot settings, indicating that models perform better in the presence of context, even without pre-training. Notably, the metrics' results during Fine-Tuning in configurations where the fallacy context is omitted (LO, NO) are significantly higher compared to the Zero-Shot and Few-Shot settings.

To address **RQ3**, I conducted a human annotation study evaluating the quality of repaired arguments generated by top-performing models. Seventeen annotators voluntarily analyzed 15 repaired arguments per model using the human evaluation metrics above-mentioned, in a controlled annotation environment to promote real-time feedback and problem-solving. The analysis of inter-rater reliability, measured using Krippendorff's α, revealed values ranging from 0.16 to 0.22 across all criteria, indicating low reliability, as shown in Table 7.2.

	Krippendorf's α	**% Agreement**	**Mean**
Relevance	0,16	55%	$4.03 \pm 0{,}68$
Suitableness	0,19	60%	$4.17 \pm 0{,}68$
Cogency	0,19	49%	$3.76 \pm 0{,}69$

Table 7.2: Human evaluation results for **RQ3** based on the annotation of 15 repaired arguments generated by LLMs and annotated by 17 human annotators.

Percentage agreements varied from 49% to 60%, with Suitableness showing the highest agreement. These results highlight the subjective nature of argument evaluation and the challenges faced in analyzing complex linguistic tasks. Despite the variability in inter-rater agreement, the LLM-generated annotations received promising scores on a 5-point scale: Relevance (4.03), Suitability (4.17), and Cogency (3.76). While these scores demonstrate strong performance in Relevance and Suitableness, the lower Cogency rating suggests an area for future improvement in enhancing the persuasiveness of repaired arguments. The demographic profile of our annotators, predominantly in the 21-30 age group and largely comprising Master's and PhD students, provided valuable academic perspectives, as shown in Figure 7.5.



(a) Age **(b)** Genders **(c)** Education level

Figure 7.5: Demographics distribution of the 17 volunteer annotators.

7.4 Error Analysis

Regarding RQ1, an analysis of Table 7.1 reveals that performance, as measured by the macro average F1 Score, varies significantly across Zero-Shot (ZS), Few-Shot (FS), and Fine-Tuning (FT) settings in both configurations. The Context Only (`CO`) configuration consistently outperforms the No Fallacy Label & Context (`NO`) configuration across most models and settings. Notably, Few-Shot (FS) and Fine-Tuning (FT) approaches generally achieve superior performance metrics compared to Zero-Shot (ZS) in both configurations. The most significant improvements are observed in the `CO` configuration, where models are provided with the fallacious context. For instance, GPT-4 demonstrates a substantial performance boost in the FS setting, achieving a macro average F1 Score of 59.15% in `CO`, compared to 37.69% in `NO` — a difference of 21.46 percentage points. Interestingly, LLaMA 3 8B exhibits the highest performance in the FT setting for both configurations, with 52.76% in `CO` and 31.06% in `NO`, demonstrating the potential of fine-tuning for this specific task. The baseline BART model, evaluated only in the FT setting, shows a notable improvement when fine-tuned with contextual information, achieving 34.92% in `CO` compared to 42.20% in `NO`. This 7.28 percentage point increase underscores the importance of context in fallacy classification for this architecture and approach. These observations challenge the assumption that additional contextual information might introduce noise. Instead, it appears to significantly enhance the models' ability to accurately classify fallacies. This finding aligns more closely with the conventional understanding that context often leads to better performance in natural language processing tasks, particularly in nuanced areas like fallacy classification. The consistent outperformance of the `CO` configuration across most models and settings emphasizes the crucial role of context in understanding and correctly identifying fallacious arguments. To further investigate the model's performance and understand the specific challenges in fallacy classification, I conducted a detailed error analysis using a confusion matrix based on the best-performing model's predictions in FS configuration, as shown in Figure 7.6. The confusion matrix results should be interpreted in light of the class imbalance in the dataset, as detailed in Table 3.8. Loaded Language shows the best performance, which is expected given its predominance. Its high accuracy is likely influenced by the wide range of training examples. Flag Waving

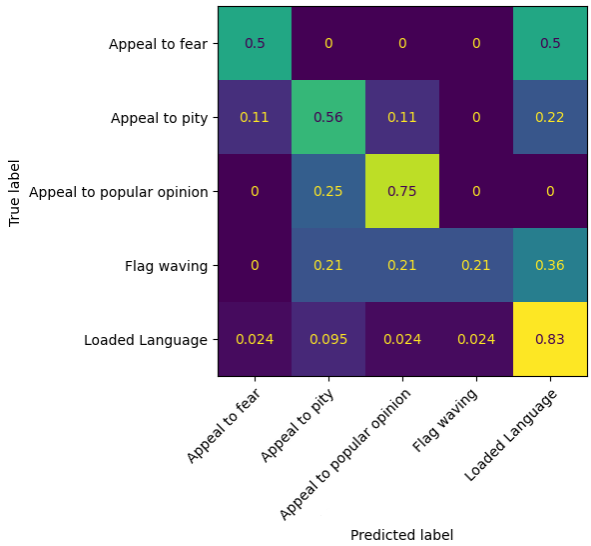


Figure 7.6: Confusion matrix of LLaMa 3 8B's performance in CO setting during Few-Shot experiments.

surprisingly underperforms despite its substantial representation. This suggests particular difficulty in capturing its unique features, possibly due to overlap with other categories. The model's moderate performance on Appeal to Fear and poor performance on Appeal to Popular Opinion align somewhat with their lower representation in the dataset. However, Appeal to Pity's weak performance is noteworthy given its relatively higher frequency, indicating that factors beyond mere class representation are affecting classification accuracy. This imbalanced distribution highlights the challenges in fallacy classification, where certain types are naturally more common. It emphasizes the need for strategies to address class imbalance, such as weighted loss functions or data augmentation for underrepresented categories, to improve overall model performance across all fallacy types.

	Context Only (CO)			No Fall. & Ctx (NO)		
Model	**ZS**	**FS**	**FT**	**ZS**	**FS**	**FT**
BART	-	-	5	-	-	5
Claude 3	7	7	-	13	9	-
Gemma 1.1 2B	8	172	8	7	4	5
Gemma 1.1 7B	71	5	14	70	99	13
GPT-3.5	19	9	10	23	6	10
GPT-4	7	5	-	9	6	-
LLaMA 3 8B	34	7	5	23	25	2
Mistral 7B	368	264	-	156	182	-
Mixtral 8x7B	193	159	12	158	112	14

Table 7.3: Distribution of labels predicted by the models across the five proposed fallacy categories in CO and NO configurations.

Further analysis reveals a nuanced qualitative discrepancy in LLM performance for fallacy classification. While LLaMA 3 achieved a notably high macro F1 Score in the Fine-Tuning (FT) setting with the context-free configuration (`NO`) as shown in Table 7.1, a closer examination of label assignment patterns, presented in Table 7.3, suggests potential limitations in the model's classification strategy. Specifically, LLaMA 3's high accuracy in the `NO` configuration was achieved using a significantly reduced set of only 2 labels. This lack of precision in label choice, while effective in terms of overall accuracy, raises questions about the model's ability to discriminate between the full spectrum of fallacy types proposed in the prompt (5 categories). In contrast, GPT-4 demonstrated a tendency to exceed the established limit of five labels. In the Zero-Shot (ZS) setting, GPT-4 predicted 9 labels in the `NO` configuration and 7 labels in the Context Only (`CO`) configuration. Similarly, in the Few-Shot (FS) setting for the `NO` configuration, GPT-4 generated 6 distinct labels. This overproduction of labels suggests a more expansive, albeit potentially less constrained, approach to fallacy classification. Notably, other models exhibited similar tendencies to stick to the five-category limit of fallacies. Google's Gemma models, particularly Gemma 1.1 2B, dramatically exceeded the limit in the FS setting for the `CO` configuration, predicting 172 distinct labels. Mistral 7B and Mixtral 8x7B also consistently produced a high number of fallacy labels across various settings, with Mistral 7B predicting up to 368 labels in the ZS setting for `CO` configuration, as shown in Table 7.4.

Subcategories for fallacy classification
Appeal to Uninformed Patriotism, Appeal to Emotion (specifically, an Appeal to Consequence), Appeal to Emotion (specifically, Appeal to Nationalism), Appeal to Emotion (contains elements of Appeal to pity and Appeal to fear), Flag waving, Loaded Language, Appeal to Emotion, Unclear / Vague Argument, Appeal to Emotions (specifically, Appeal to pity), Appeal to Misleading Statistics, Appeal to Expertise, Appeal to fear and Appeal to pity, Appeal to Emotion (specifically, Appeal to Sentiment), Appeal to Personal Privilege or Experience (also known as "Appeal to Authority" if the speaker is an actual expert), Appeal to Tradition, Appeal to Consequences, Appeal to Personal Experience, Ambiguity or Vagueness, Appeal to Hypothetical, Appeal to Unverified Experience, Appeal to Emotion (specifically, an Appeal to Concern), Appeal to fear, Appeal to Emotion (specifically, Appeal to fear and Negative Stereotyping), Appeal to Ancedote, Appeal to Popular Opinion, Appeal to Emotion (specifically, Appeal to Patriotism), Appeal to Emotion (specifically, Appeal to pity and Appeal to fear), Appeal to Emotion (specifically, Appeal to Consequences), Appeal to pity, Appeal to Confidently Held Belief, Appeal to Self-Interest (or Appeal to Personal Benefit), Loaded Language, Appeal to Emotion, Appeal to Unintended Consequences, Appeal to Misleading Numbers, Appeal to Authority, Appeal to Morality, Appeal to Authority or Appeal to Expertise, Appeal to Emotion (includes Appeal to pity and Appeal to fear), Appeal to pity, Appeal to Values (specifically, an Appeal to Morality), Appeal to Emotion (Flag waving), Appeal to Authority or Appeal to Expertise, Appeal to fear, Appeal to Emotion (specifically, Appeal to pity and Appeal to Disgust), Appeal to Time, Appeal to Emotion (specifically, Appeal to pity, Loaded Language, Appeal to Self-righteousness or Pride, Appeal to Ignorance, Appeal to pity and Appeal to Authority, Loaded Language, Ad Hominem, specifically the subtype "Accusation of Inconsistency", Appeal to Novelty or Ignorance, Appeal to Anecdote, Appeal to Emotion (specifically, Appeal to Hope and Self-Interest), Appeal to fear and Appeal to Authority, Equivocation, Appeal to Emotion (specifically, Appeal to fear), Appeal to Identity and Populism, Appeal to Authority, Appeal to Slippery Slope, Appeal to Pity and Statistics without Contextn, Appeal to Past Failure, Appeal to Anecdote or Appeal to Personal Experience, Appeal to Anecdote (a type of Appeal to Experience), Appeal to fear and Loaded Language, Appeal to Ideal, Appeal to Emotion (specifically, Appeal to fear and Appeal to pity), Appeal to Patriotism (similar to Flag waving), Appeal to Anecdote, Appeal to Emotion (specifically, Appeal to Action), Appeal to Circumstances (a form of Appeal to Emotion), Appeal to Expert Opinion, Appeal to Popularity (or Appeal to the People), Appeal to Consequences (a type of Appeal to Emotion), Appeal to pity and Class Stereotyping, Appeal to Consequences (a type of Appeal to fear), Appeal to Emotion (specifically, an Appeal to pity), Appeal to Popular Opinion (or Appeal to Authority), Appeal to Emotion (specifically, Appeal to Disgust), Appeal to Consequences (specifically, an unstated consequentialist argument), Appeal to Authority or Expertise, Appeal to False Cause, Appeal to Popular Opinion (or Demagoguery), Appeal to Emotion (specifically, Appeal to pity), Appeal to Identity/Demographic Appeal, Appeal to Popular Opinion (or Appeal to the People), Appeal to Popular Opinion, Appeal to Statistics, Appeal to Changed Circumstances (or Hypocrisy), Appeal to Ethics (specifically, Appeal to Morality), Appeal to Emotion (Specifically, Appeal to pity), Unsupported Claims and Vague Quantifiers

Table 7.4: Examples of *over-predicted* fallacy labels by LLMs, specifically Mistral 7B in Zero-Shot setting.

These unexpected divergences from the suggested category set highlight significant challenges in constraining language models to a predefined classification schema.

Such observations underscore a fundamental challenge in aligning LLMs' internal representations with externally defined classification schemas. Despite explicit instructions to select labels from a specific set of five categories, many models demonstrated a tendency to generate a broader, and sometimes vastly expanded, range of labels. This phenomenon points to a potential mismatch between the LLMs' learned representations of fallacies and the prescribed taxonomic structure of the task. This divergence between the expected and the actual model behavior invites deeper investigation into the internal mechanisms governing label assignment within LLMs, particularly when constrained by a closed set of possibilities. Key questions emerge regarding the models' ability to adhere to explicit classification boundaries and the nature of their learned representations of logical fallacies. For instance, how do these models interpret and process the constraints provided in the prompts, and what factors influence their propensity to generate labels beyond the specified set? Furthermore, the dichotomy between high quantitative performance (as measured by F1 scores) and qualitative inconsistencies in label assignment underscores the importance of a variety of evaluation metrics in assessing LLM performance on complex linguistic tasks. This finding suggests that traditional performance metrics may not fully capture the nuances of model behavior in tasks requiring fine-grained semantic distinctions, such as fallacy classification. Consequently, it becomes crucial to develop more comprehensive evaluation frameworks that can account for both quantitative accuracy and qualitative adherence to task specifications.

In relation to **RQ2**, the error analysis reveals significant variations in model performance across different settings and configurations. Focusing first on the Zero-Shot (ZS) setting, we observe that GPT-4 demonstrates superior performance in the `LC` and `LO` configurations, as evidenced by its high BERTScore and IOU-F1 values (Figure 7.2). However, this performance comes with a drawback: GPT-4 exhibits a high percentage of responses that do not match the prompted instructions, particularly in the `LC` (91.43%) and `LO` (74.86%) configurations, as showed in Table 7.6. The 'Not Matched' percentage in Table 7.6 represents LLMs' responses that failed to follow the specified output schema (see 7.1.1). For instance, GPT-4 often omitted the required "- Fixed Argument:" prefix, instead using variations like "Argument: [...]" or presenting the repaired argument without any prefix. This inconsistency, while not necessarily indicative of incorrect content, violates the strict formatting requirements of this evaluation. Some examples are shown in Table 7.5 where GPT-4 responds directly without following the prompt's instructions.

Interestingly, other models display varying degrees of adherence to instructions across different configurations. In the ZS setting, Claude and Gemma 1.1 2B show perfect adherence (0% Not Matched) for all configurations except `NO` for Claude (0.37%). Mixtral 8x7B, despite its generally lower adherence, demonstrates better consistency across configurations compared to GPT-4. Transitioning to the Few-Shot (FS) setting, it is possible to observe a substantial shift in model behaviors. GPT-4's adherence issues are exacerbated, with 100% Not Matched for `LC` and 85.03% for `LO`. Surprisingly, LLaMA

Prompt	**Response**
"Given a fallacious argument 'The position of the United States has been that this build-up, in the words of the president, has been foolish.' labeled as: 'Loaded Language', [...] Output of the answer in the following structure: - Fixed Argument:"	"The United States has taken the stance that this build-up, as stated by the president, may not have been the most strategic decision."
"Given a fallacious argument 'We were threatened by a veto if we passed a dollar and a quarter - it's extremely difficult with the great power that the president does to pass any bill when the president is opposed to it.' labeled as: 'Loaded Language', [...] Output of the answer in the following structure: - Fixed Argument:"	"The president has expressed opposition to our proposed bill and has even threatened a veto. Given the significant influence the president holds, passing a bill under these circumstances is a challenging task."
"Given a fallacious argument 'And to make sure our society is as hopeful as it possibly can be, every single child in America must be educated. I mean every child.' labeled as: 'Flag waving', [...] Output of the answer in the following structure: - Fixed Argument:"	"To ensure the maximum potential of our society, it is crucial that we prioritize the education of every child in America."

Table 7.5: Examples of GPT-4 responses without following the instructions prompted.

3 8B shows a significant increase in non-adherence for the CO configuration (55.56%), while maintaining better adherence in others. Claude improves its performance in NO and LO but shows a notable increase in non-adherence for LC (3.75%). The Fine-Tuning results, where available, present another layer of complexity. Models like Gemma 1.1 2B and Mixtral 8x7B show improvements in certain configurations while maintaining perfect alignment in others.

	Zero-Shot				*Few-Shot*				*Fine-Tuning*			
Model	CO	NO	LC	LO	CO	NO	LC	LO	CO	NO	LC	LO
Claude	0%	0,37%	0,13%	0,18%	0%	0%	3,75%	0,37%	-	-	-	-
Gemma 1.1 2B	0%	0%	0%	0%	0%	0%	0%	0%	2,67%	1,92%	0%	0%
Gemma 1.1 7B	0,27%	0%	0%	0,18%	0%	0%	0%	0%	0%	0%	0%	0%
GPT-3.5	3,21%	0,54%	0%	0,18%	3,35%	3,08%	99,06%	1,85%	5,33%	1,92%	2,67%	1,92%
GPT-4	0,67%	0%	91,43%	74,86%	1,47%	0,27%	100%	85,03%	-	-	-	-
LLaMA 3 8B	0%	0%	0,40%	0,74%	55,56%	0,54%	12,32%	0,18%	0%	0%	0%	1,92%
Mistral 7B	0,80%	1,47%	0,13%	0,18%	0%	0,94%	43,78%	0%	-	-	-	-
Mixtral 8x7B	17,40%	12,99%	1,07%	8,50%	4,55%	5,89%	2,68%	0,92%	0%	3,85%	0%	1,92%

Table 7.6: Results in percentage of models that did not match the instructed prompt in all the configurations and settings.

Further analysis was conducted to examine the impact of including fallacy labels in the prompts given to the LLMs. Table 7.7 illustrates the changes in performance metrics when transitioning from configurations without labels (CO, NO) to those with labels (LC, LO). This comparison yields several noteworthy insights across different settings. In the Zero-Shot (ZS) setting, the inclusion of fallacy labels shows a context-dependent impact. When context is present, transitioning from CO to LC configurations yields positive results, with improvements in both BERTScore (+2.79%) and IOU-F1 (+0.80%). However, without context, moving from NO to LO configurations produces a negative trend, with decreases in both BERTScore (-3.07%) and IOU-F1 (-6.06%). This dichotomy

Metric	Config.	Zero-Shot	Few-Shot	Fine-Tuning
BERTScore	CO	$0,6932_{GPT-4}$	$0,7063_{Claude}$	$0,9331_{Llama3}$
	LC	$0,7125_{GPT-4}$	$0,7753_{Claude}$	$0,8780_{Llama3}$
Avg		*+2,79%*	*+9,77%*	*-5,91%*
IOU-F1	CO	$0,8936_{GPT-4}$	$0,8917_{Claude}$	$0,9723_{Llama3}$
	LC	$0,9008_{GPT-4}$	$0,9180_{Claude}$	$0,9723_{Llama3}$
Avg		*+0,80%*	*+2,95%*	*+0,00%*
BERTScore	NO	$0,6199_{GPT-3.5}$	$0,6864_{GPT-3.5}$	$0,9560_{GPT-3.5}$
	LO	$0,6009_{GPT-4}$	$0,7180_{Gemma7B}$	$0,9693_{Llama3}$
Avg		*-3,07%*	*+4,61%*	*+1,38%*
IOU-F1	NO	$0,7863_{GPT-3.5}$	$0,8082_{GPT-3.5}$	$0,9562_{Llama3}$
	LO	$0,7387_{GPT-4}$	$0,8290_{Gemma7B}$	$0,9777_{Llama3}$
Avg		*-6,06%*	*+2,57%*	*+2,25%*

Table 7.7: Impact of including fallacy labels in the prompt on evaluation metrics based on the best model for each setting: Zero-Shot, Few-Shot, and Fine-Tuning.

suggests that in ZS scenarios, the effectiveness of fallacy labels is highly dependent on the availability of contextual information. The Few-Shot (FS) setting presents a more consistent picture. Across all configurations, the inclusion of fallacy labels leads to improvements in model performance. With context, BERTScore shows a substantial increase of 9.77%, while IOU-F1 improves by 2.95%. Even without context, both metrics see positive changes, with BERTScore increasing by 4.61% and IOU-F1 by 2.57%. These results indicate that in FS scenarios, fallacy labels consistently enhance model performance, with a more pronounced effect when context is available. The Fine-Tuning (FT) setting, however, presents a more nuanced picture. With context, including fallacy labels, leads to a decrease in BERTScore (-5.91%), while IOU-F1 remains unchanged. Conversely, without context, both metrics show slight improvements (BERTScore: +1.38%, IOU-F1: +2.25%). This mixed outcome suggests that for fine-tuned models, the benefit of including fallacy labels is less clear-cut and may depend on specific configurations. Examining model performance across settings reveals interesting patterns. In the ZS setting, GPT-4 consistently performs best, except for the NO configuration where GPT-3.5 excels. The FS setting sees Claude dominating in configurations with context, while GPT-3.5 and Gemma7B perform best in NO and LO configurations, respectively. In the FT setting, Llama3 shows superior performance across all configurations, with GPT-3.5 matching it in the NO configuration. These findings highlight the complex relationship between prompt design, model architecture, and training methods in the task of repairing fallacious arguments. The generally positive impact of including fallacy labels, especially in Few-Shot settings, suggests that explicit identification of fallacy types can guide models toward more accurate repairs. However, the varied results across settings underscore the challenge of identifying a universally optimal approach. The mixed outcomes, particularly in the Fine-Tuning setting, indicate that the relationship between fallacy labeling and argument repair quality is not straightforward. This complexity may be attributed to the inherent difficulty and novelty of the task, which

requires LLMs to not only understand political fallacies but also to generate reliable, corrected arguments.

Thus, the error analysis of **RQ3** focused on inter-rater reliability in evaluating LLM-generated outputs. I observed a high percentage of agreement among annotators, suggesting LLMs generally produce fitting and relevant content. However, this analysis also revealed significant subjectivity in evaluations, with annotators often reaching similar judgments through different reasoning processes. The presence of a discrete number of '% Not Matched' responses further confuses the picture. While high agreement rates are positive, these instances of non-matching indicate that LLM outputs still require improvement in coherence and overall quality. This suggests that while LLMs can produce generally acceptable outputs, there remain cases where the generated content falls short of human expectations or fails to fully address the nuances of the original fallacious argument.

7.5 Summary

This chapter addresses the complex challenge of repairing fallacious arguments in political debates. Drawing on the FallacyFix dataset, this study implements diverse large language models to achieve two goals simultaneously: fallacy type classification and argument correction. Through a comprehensive analysis using tailored prompts, the research reveals promising results in Fine-Tuning settings, while also highlighting the limitations of Zero-Shot and Few-Shot approaches, particularly in adhering to prompt instructions. Thus, to validate the practical feasibility of this approach, an extensive user study was conducted, confirming the relevance, suitability, and cogency of the repaired arguments. By advancing both the technical aspects of fallacy detection and repair and offering insights into political discourse analysis, this research has significant implications for enhancing the quality of public debate and political communication.

Chapter 8

Conclusion and Prospectives

As previously introduced, the analysis of argumentation, particularly within the context of political debates, presents a significant challenge in the field of Natural Language Processing. Unlike other areas of NLP, the study of political debates and the in-depth analysis of related information have only recently become subjects of scientific research. Consequently, the selection, evaluation, and application of this information as relevant evidence for the debate analysis process prove to be complex and labor-intensive tasks. This complexity underscores the need for systems capable of (semi)automatically assisting in the processing of large volumes of data. Political debates offer a rich corpus of textual data for argumentation analysis. In these contexts, candidates employ arguments to justify past actions, present future plans, and critique opponents' positions. Despite the existence of numerous Argument Mining approaches and annotated corpora, few studies have applied to political texts, and very few have yet comprehensively addressed the problem of extracting complete argumentative structures from such texts. This thesis focused on two primary issues:

1. The scarcity of adequate textual data: To address this challenge, two novel datasets were proposed: ElecDeb60to20 and FallacyFix. The first is dedicated to argumentative analysis, while the second focuses on the elaboration of specific argumentative elements, with particular attention to fallacies.

2. The argumentative analysis of political debates: After collecting and annotating data across various argumentative levels (components, relations, and fallacies), this thesis proposed several approaches for processing and analyzing political debates. These approaches aim to automate the extraction of argumentative components (such as premises and evidence), the identification of their relations (attack or support), as well as the detection and classification of fallacies.

In summary, the research conducted in the context of this thesis demonstrates the application and development of Argument Mining methods for the political domain, particularly focusing on political debates. As this field is still evolving, and to foster future research in the area of Argument Mining on political discourse, two novel datasets have been made available to the research community, along with the source code of the experiments. The contributions outlined in this thesis serve as a significant foundation,

aiming to inspire and encourage the research community to further develop these ideas and promote the utilization and refinement of the presented datasets and approaches.

To address the research questions presented in Chapter 1, this thesis offers several novel solutions, resulting in the following key contributions:

Contribution 1 — Updated Argument Mining Tool for Political Debates with New Dataset, Views, and Argument Mining Pipeline DISPUTool 2.0 addresses critical gaps in the analysis of political debates by providing a comprehensive, integrated solution to several persistent challenges in Argument Mining. This upgraded version tackles the current problem of fragmented analysis tools by integrating improved argumentative component identification, relationship mapping, Named Entity Recognition, and graph overviews into a single platform. This holistic approach solves the issue of researchers having to use multiple, often incompatible tools for debate transcript analysis. By bridging computer science and humanities, DISPUTool 2.0 offers a solution to the interdisciplinary divide that has hindered progress in the computational analysis of political argumentation. The inclusion of graph overviews provides a novel solution to the challenge of visualizing complex argumentative structures, enabling researchers to more easily identify patterns and relationships within debates. Furthermore, by incorporating the updated dataset, the tool addresses the persistent issue of outdated training data in rapidly evolving political landscapes. DISPUTool 2.0 thus fills a crucial gap in the study of political argumentation, offering researchers a sophisticated, all-in-one solution for exploring and analyzing debate content, potentially revolutionizing our understanding of political discourse and opening new avenues for automated fact-checking and argument quality assessment.

Contribution 2 — Domain-Specific Approach to Apply Argument Mining for Political Debate Analysis This contribution addresses the challenge of effectively mining arguments from complex political debates by proposing a domain-specific definition of Argument Mining tasks. It tackles the issue of generic Argument Mining approaches failing to capture the nuances of political discourse. The novel two-stage Argument Mining pipeline integrates both component detection and relation classification, solving the problem of disjointed argument analysis. By employing advanced neural networks, it overcomes the limitations of traditional methods in identifying subtle argument components within debate text. The second stage addresses the critical issue of understanding argument structure by predicting relationships between components. This integrated approach demonstrates significant practical effectiveness, surpassing standard baselines with average F1-scores of 47% for component detection and 69% for relation prediction. These improvements over existing methods provide a solution to the persistent problem of low accuracy in automated political argument analysis. The results, validated through extensive linguistic analysis and error investigation, not only confirm the feasibility of accurate Argument Mining in political debates but also offer a robust framework for future research.

Contribution 3 — Updated ElecDeb60to20 Dataset with Argumentative and Fallacious Annotations This contribution directly addresses the critical issue of outdated and limited datasets in the field of Argument Mining for political debates. Following the guidelines in [63], I updated the ElecDeb60to16 dataset, continuing to solve the problem of data scarcity and lack of contemporary examples in existing resources. This updated dataset serves as a crucial solution for training and evaluating classifiers in the supervised classification approach to AM in political debates. The original dataset, comprising 54,640 components (29,004 claims and 25,635 premises), 25,012 relations (21,289 support and 3,723 attack), and 1,628 fallacies from 41 debates, including 8 vice-presidential debates from 1960 to 2016, has been significantly expanded. This expansion tackles the issue of limited data diversity by incorporating the final debates from the 2020 U.S. presidential election between Biden and Trump, contributing 3 new debates, 1,038 components, 512 relations, and 232 fallacious arguments to the corpus. This addition not only increases the dataset's size but also addresses the critical need for up-to-date examples that reflect current political discourse and rhetorical strategies. To the best of my knowledge, this is now the largest annotated dataset within the Argument Mining field on political debates, offering a solution to the long-standing problem of insufficient data for robust model training and evaluation in political debates. The updated ElecDeb60to20 dataset stands as an even more comprehensive and unique resource, addressing the challenge of limited temporal scope in existing datasets by capturing the evolution of political debates over six decades of U.S. presidential debates. This expanded dataset offers researchers a rich corpus for exploring various aspects of political argumentation, from basic argument structures to complex fallacious reasoning patterns. It provides a potential solution to the challenge of developing more sophisticated AM models that can handle the nuances and complexities of real-world political debates.

Contribution 4 — Modelling Detection and Classification of Fallacies in Political Debates This contribution addresses the critical challenge of automatically detecting and classifying fallacies in political debates, a task that has long eluded computational approaches due to the complexity and nuance of political discourse. To tackle this issue, I developed and evaluated advanced Transformer-based models, directly addressing the limitations of existing fallacy detection systems. Initially, I enhanced the Longformer model [13] by incorporating argumentative feature labels (component and relation), solving the problem of context loss in long political texts and significantly improving the F1 Score from 0.61 to 0.84 in 6 multi-class classification. This substantial improvement demonstrates a solution to the persistent issue of low accuracy in automated fallacy detection and the need for considering the context around a specific text. Building on this success, I implemented MultiFusion BERT, a novel model that addresses the challenge of integrating multiple linguistic features by combining argument components, relations, and Part-of-Speech tags during the training phase. This innovative approach achieved an average F1 Score of 0.74 in detecting and classifying fallacies, improving upon the starting approach's performance of 0.72 F1 Score and

outperforming existing methods. This advancement offers a solution to the problem of accurately identifying fallacy boundaries within complex argumentative structures. Extensive error analysis revealed the model's robustness in handling nuanced political discourse while identifying areas for improvement, providing a roadmap for future enhancements in fallacy detection systems. Moreover, this work addresses the urgent need to scrutinize political arguments for sound reasoning, a critical issue in an era of increasing misinformation and manipulative rhetoric. By providing insights into the strategic use of fallacies in influencing opinion and diverting attention, it offers a potential solution to the challenge of automated fact-checking and argument quality assessment in political debates. The emphasis on the interplay between rhetoric, philosophy, and political communication [167, 165] provides a comprehensive framework for understanding and countering fallacious reasoning in political discourse, potentially improving the quality of public debate and decision-making.

Contribution 5 — Creation of a New Dataset of Annotated Repaired Fallacies To tackle the critical gap in resources for fallacy correction in political discourse, FallacyFix: A Repaired Fallacies Dataset has been created. This novel gold standard dataset directly addresses the fundamental challenge of repairing fallacious reasoning in political debate arguments, a problem that has hindered progress in automated argument improvement systems. Built upon the ElecDeb60to20-fallacy dataset, FallacyFix offers a solution to the lack of comprehensive, real-world examples of fallacy repair in political contexts. FallacyFix focuses on systematically repairing fallacious arguments, particularly those involving Appeal to Emotion and Appeal to Authority, addressing the specific need for resources targeting common rhetorical fallacies in political debates. The creation process involved expert annotators employing a straightforward methodology of removal, subtraction, and simplification, considering the full context of each fallacious argument. This approach ensures high-quality annotations while managing the complexity inherent in fallacy repair, solving the problem of oversimplified or context-ignorant fallacy correction methods. To validate the consistency and quality of the repairs, a set of 100 fallacious examples randomly extracted from the FallacyFix dataset was annotated to assess agreement between two generated texts. Using BERTScore [173] and BERT [36] embeddings comparison yielded scores of 0.94 ± 0.06 and 0.98 ± 0.03, respectively. BERTScore was chosen as the most appropriate metric due to the current absence of standard methods for comparing repaired fallacious arguments. These high scores demonstrate the reliability and consistency of the dataset, addressing the crucial need for dependable resources in fallacy repair research. The resulting dataset is available in two versions: a standard version with 747 repaired annotations. At the moment, FallacyFix represents a significant contribution to the field of Argument Mining, offering researchers a unique resource for developing and evaluating techniques to analyze and repair fallacious reasoning in political debates.

8.1 Limitations

While these contributions have made significant strides in argument mining for political debates, it is important to acknowledge the limitations that emerged throughout the study. These constraints not only shape the interpretation of my results but also point toward promising avenues for future research in this field. This thesis, although advancing the understanding of argumentation in political discourse, faced several challenges. The use of advanced transformer models and large language models, while powerful, introduced its own set of limitations. The trade-off between the performance of proprietary models and the resource demands of open-source alternatives presented a significant consideration for practical implementation. Moreover, the focus on US political debates in our dataset, while providing rich contextual data, inevitably restricted the model's applicability to English-language contexts, highlighting the need for more diverse, multilingual datasets in future studies. Data imbalances, particularly the underrepresentation of certain argumentative categories like "Slogan," affected the model's performance and generalizability. This imbalance points to the importance of developing more comprehensive and evenly distributed datasets for training robust argument mining models. The complexity of evaluating arguments extracted from fallacies against generated ones highlighted the need for standardized metrics in this domain. The subjective nature of argument assessment, particularly evident in human annotation processes, underscored the difficulties in achieving consistent and reliable evaluations. This subjectivity became even more pronounced when attempting automated assessments, revealing a critical area for methodological improvement. Additionally, the modest performance gains observed, despite significant increases in model complexity, suggest that future research should explore more efficient architectural designs and training strategies. The resource-intensive nature of our approach, requiring high-end hardware, raises questions about the accessibility and scalability of these methods in resource-constrained environments.

8.2 Ethical considerations

While enhancing the objective of public understanding and resilience against misinformation holds significant potential for societal benefit, it is crucial to address the ethical implications and potential unintended consequences. The application of AI in analyzing political discourse presents both opportunities and challenges, including the risk of amplifying existing biases, overlooking underrepresented perspectives, and the potential misuse of these technologies to create misleading content. Challenges such as ensuring reliability, mitigating dataset biases, preventing misuse, and addressing privacy concerns are crucial. To address these issues, an approach could be implementing rigorous human expert review processes, providing clear disclaimers for AI-generated content, expanding dataset diversity, developing strict ethical guidelines, and creating safeguards against misuse. By balancing the potential benefits with careful consideration of risks, we strive to advance this field ethically and responsibly. This approach

could emphasize adaptability, transparency, and ongoing ethical reflection, recognizing that the implications of AI in political discourse extend beyond technical considerations to fundamental aspects of democratic participation and information integrity. Through these measures, we aim to contribute positively to the understanding and quality of political discourse while safeguarding the integrity of democratic processes and public trust in political institutions.

8.3 Prospectives

Although this thesis has established significant conceptual building blocks, it also sheds light on several possible directions for future research and potential enhancements in the field.

Future perspectives in the field of political argumentation and Argument Mining are articulated on several innovative fronts, promising to significantly enrich our understanding and analysis of political discourse. As highlighted by [166], the development of larger and more diverse corpora, enriched with multi-level argumentative annotations, is fundamental for an in-depth analysis of argumentative structures. The expansion of corpora to political debates from different nations would allow comparative analyses, revealing similarities and differences in argumentative strategies across cultural contexts, in line with Tindale's [153] view on the importance of context in argumentation. [101] emphasized the importance of developing more extensive corpora focused on counter-argumentation techniques and fallacy repair. As we expand corpora to include political debates from different nations and cultures, it is crucial to develop culturally sensitive annotation guidelines and models. Argumentation strategies and norms can vary significantly across cultural contexts, influenced by factors such as history, religion, social structures, and political systems [64]. These differences can manifest in various aspects of argumentation, such as the use of rhetorical devices, the emphasis on certain types of evidence, and the acceptability of emotional appeals [113]. Failing to account for these cultural differences can lead to biased or ineffective AM models that misinterpret or overlook important aspects of argumentation. To address this challenge, researchers must involve domain experts from diverse cultural backgrounds in the development of annotation guidelines and models [77]. These experts can provide valuable insights into the nuances of argumentation within their respective cultures, helping to create guidelines that capture a wide range of argumentative patterns and strategies. Moreover, it is essential to develop flexible annotation schemes that can accommodate cultural variations while maintaining a consistent framework for analysis [90]. By incorporating cultural sensitivity into the development of AM models, we can improve their accuracy, generalizability, and ability to provide meaningful insights into the diversity of political discourse worldwide. This approach not

only enhances the effectiveness of AM tools but also promotes a more inclusive and respectful understanding of argumentation across cultures.

Another promising direction in the AM field is the improvement of multimodal models that integrate text, audio, and video. The advancements in multimodal Argument Mining shows promise but also reveals complexities. While integrating audio and text modalities showed improvements in certain tasks, such as detecting attack relations in imbalanced datasets [106] and outperforming text-only models in argumentative fallacy classification [107], the gains were not consistently substantial across all tasks. This variability suggests that the relationship between textual and audio features in argumentative discourse is complex and task-dependent. Current limitations, including annotation schemes that may overlook crucial acoustic cues [105] and the need for more sophisticated fusion strategies, present opportunities for future research. As demonstrated by the M-Arg dataset [114], incorporating audio features can provide added value, particularly in political debates. The association and synchronization of text, audio, and video of political debates would allow simultaneous analyses on multiple levels. For example, the analysis of vocal tone during the exposition of facts or premises, as suggested by [10], could reveal subtle emotional nuances not evident in the text alone. Similarly, the study of facial expressions and posture, based on the work of [46], could provide valuable clues about the speaker's emotional state and sincerity. As highlighted by [31], the multimodal approach is particularly effective in identifying and analyzing complex affective states, crucial in the context of political argumentation.

Furthermore, a promising practical application of these advancements could be the updating and expansion of existing tools such as DispuTool, an AM tool specifically designed for political debates. [90] emphasized the importance of developing user-friendly tools for argumentation analysis, making AM accessible to a wider audience. The integration of new methodologies could enrich DispuTool with advanced features, including: comparative exploration of international political debates, advanced visualizations of argumentative structures [133], multimodal representations of arguments, detailed explanations on the presence of fallacies or fake news, and high-performance model for extracting argumentative components and relations. These improvements would not only benefit the scientific community but would position DispuTool as a valuable educational tool. As highlighted by [87], tools that allow students to analyze and construct arguments interactively can significantly improve their critical thinking skills. DispuTool could be designed to adapt to different educational levels, following the concept of "scaffolding" in education [155].

These short-term advancements in argument mining tools like DispuTool not only offer immediate benefits but also pave the way for more ambitious research goals. As we progress from tool-specific improvements to broader horizons, the evolution of argument mining calls for a more comprehensive approach. This shift necessitates expanding our research scope to address the complex challenges of analyzing political argumentation at scale, setting the stage for interdisciplinary collaboration in the field.

To accelerate progress in this domain, fostering interdisciplinary research teams and projects is crucial. By bringing together experts from diverse disciplines such as Linguistics, Logic, Philosophy, Psychology, and Sociology, we can unlock new insights and develop more comprehensive models for analyzing political argumentation, as emphasized by [71]. By collaborating across disciplinary boundaries, researchers can develop innovative approaches, share knowledge, and tackle the challenges of AM from multiple angles. This cross-pollination of ideas can lead to breakthroughs in areas such as multimodal analysis, cultural adaptability, and ethical considerations. Each field contributes unique perspectives and methodologies that, when combined, can lead to a more nuanced understanding of the complexities inherent in political discourse. For instance, linguists can provide valuable insights into the structure and meaning of arguments, while psychologists can shed light on the emotional and cognitive aspects of persuasion. Philosophers can contribute to the development of formal models of argumentation, and sociologists can offer insights into the social and cultural factors that shape political debates. Ultimately, fostering such interdisciplinary collaboration is essential for advancing the field of AM and unlocking its full potential in understanding and analyzing political argumentation.

In conclusion, the practical application of these new theories and methods as user-friendly tools contributes to a significant move towards making argumentation analysis accessible to a wider audience. As argued by Habernal et al. [60], Argument Mining has the potential to improve media literacy and counter disinformation. Despite future challenges, the progress outlined offers a solid foundation for the future of AM and political argumentation analysis, paving the way for more informed and conscious civic engagement.

Bibliography

[1] P. C. Abrami, R. M. Bernard, E. Borokhovski, D. I. Waddington, C. A. Wade, and T. Persson. "Strategies for teaching students to think critically: A meta-analysis". In: *Review of educational research* 85.2 (2015), pp. 275–314.

[2] A. Addawood and M. Bashir. ""What is your evidence?" A study of controversial topics on social media". In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. 2016, pp. 1–11.

[3] AI@Meta. *The LLaMA 3 Foundational Language Model*. https://ai.meta.com/blog/meta-llama-3/. Accessed: 31 May 2024. 2024.

[4] A. Akbik, D. Blythe, and R. Vollgraf. "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by E. M. Bender, L. Derczynski, and P. Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1638–1649.

[5] T. Alhindi, T. Chakrabarty, E. Musi, and S. Muresan. "Multitask Instruction-based Prompting for Fallacy Recognition". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 8172–8187.

[6] T. Alhindi, S. Muresan, and P. Nakov. "Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition". In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting, Aug. 2024, pp. 12323–12334.

[7] Anthropic. *Introducing the Claude 3 Family of AI Models*. https://www.anthropic.com/news/claude-3-family. Accessed: 31 May 2024. 2023.

[8] R. Artstein and M. Poesio. "Survey Article: Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4 (2008), pp. 555–596.

[9] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. Simari, M. Thimm, and S. Villata. "Towards Artificial Argumentation". In: *AI Magazine* 38.3 (Oct. 2017), pp. 25–36.

[10] T. Bänziger and K. R. Scherer. "The role of intonation in emotional expressions". In: *Speech Communication* 46.3-4 (2005), pp. 252–267.

[11] P. Baroni, M. Caminada, and M. Giacomin. "An introduction to argumentation semantics". In: *The knowledge engineering review* 26.4 (2011), pp. 365–410.

[12] 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.

[13] I. Beltagy, M. E. Peters, and A. Cohan. "Longformer: The Long-Document Transformer". In: *CoRR* abs/2004.05150 (2020).

[14] T. J. Bench-Capon and P. E. Dunne. "Argumentation in artificial intelligence". In: *Artificial intelligence* 171.10-15 (2007), pp. 619–641.

[15] S. Bird and E. Loper. "NLTK: The Natural Language Toolkit". In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 214–217.

[16] L. Bode and E. Vraga. "In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media". In: *Journal of Communication* 65 (Aug. 2015).

[17] F. Boltužić and J. Šnajder. "Back up your Stance: Recognizing Arguments in Online Discussions". In: *Proceedings of the First Workshop on Argumentation Mining*. Ed. by N. Green, K. Ashley, D. Litman, C. Reed, and V. Walker. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 49–58.

[18] M. Boudry, F. Paglieri, and M. Pigliucci. "The fake, the flimsy, and the fallacious: Demarcating arguments in real life". In: *Argumentation* 29 (2015), pp. 431–456.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners". In: *CoRR* abs/2005.14165 (2020).

[20] 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.

[21] E. Cabrio and S. Villata. "Five Years of Argument Mining: a Data-driven Analysis". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 5427–5433.

[22] T. Chakrabarty, C. Hidey, S. Muresan, K. McKeown, and A. Hwang. "AMPERSAND: Argument Mining for PERSuAsive oNline Discussions". In: *CoRR* abs/2004.14677 (2020).

[23] G. Chen, L. Cheng, L. A. Tuan, and L. Bing. *Exploring the Potential of Large Language Models in Computational Argumentation*. 2024.

[24] A. Chernodub, O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann, and A. Panchenko. "Targer: Neural argument mining at your fingertips". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2019, pp. 195–200.

[25] C.-H. Chiang and H.-y. Lee. "Merging Facts, Crafting Fallacies: Evaluating the Contradictory Nature of Aggregated Factual Claims in Long-Form Generations". In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 2734–2751.

[26] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". 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. 1724–1734.

[27] N. Chomsky. *On nature and language*. Cambridge University Press, 2002.

[28] K. Clark, M. Luong, Q. V. Le, and C. D. Manning. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *CoRR* abs/2003.10555 (2020).

[29] M.-A. Clinciu, A. Eshghi, and H. Hastie. "A Study of Automatic Metrics for the Evaluation of Natural Language Explanations". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by P. Merlo, J. Tiedemann, and R. Tsarfaty. Online: Association for Computational Linguistics, Apr. 2021, pp. 2376–2387.

[30] M. Cohen and E. Nagel. *An Introduction to Logic*. Routledge Paperbacks. Routledge & Kegan Paul, 1966. ISBN: 9780367376239.

[31] S. K. D'Mello and J. Kory. "A Review and Meta-Analysis of Multimodal Affect Detection Systems". In: *ACM Computing Surveys (CSUR)* 47.3 (2015), pp. 1–36.

[32] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1377–1414.

[33] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. D. Pietro, and P. Nakov. "A Survey on Computational Propaganda Detection". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by C. Bessiere. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 4826–4832.

[34] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov. "Fine-Grained Analysis of Propaganda in News Article". 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. 5636–5646.

[35] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov. "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2978–2988.

[36] 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, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[37] J. N. Druckman. "The Power of Television Images: The First Kennedy-Nixon Debate Revisited". In: *The Journal of Politics* 65.2 (2003), pp. 559–571.

[38] R. Duthie and K. Budzynska. "A Deep Modular RNN Approach for Ethos Mining". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 2018, pp. 4041–4047.

[39] R. Duthie, K. Budzynska, and C. Reed. "Mining Ethos in Political Debate". In: *Computational Models of Argument - Proceedings of COMMA 2016, Potsdam, Germany, 12-16 September, 2016*. Vol. 287. Frontiers in Artificial Intelligence and Applications. 2016, pp. 299–310.

[40] F. H. V. Eemeren. "Fallacies". In: *Critical concepts in argumentation theory* (2001), pp. 135–164.

[41] F. H. V. Eemeren. *Strategic Maneuvering in Argumentative Discourse. Extending the Pragma-Dialectical Theory of Argumentation*. Amsterdam-Philadelphia: John Benjamins, 2010.

[42] F. H. V. Eemeren and R. Grootendorst. *Argumentation, Communication, and Fallacies a Pragma-Dialectical Perspective*. Routledge, 1992.

[43] F. H. V. Eemeren and R. Grootendorst. "Fallacies in Pragma-Dialectical Perspective". In: *Argumentation* 1.3 (1987), pp. 283–301.

[44] F. H. v. Eemeren and R. Grootendorst. *A Systematic Theory of Argumentation: The pragma-dialectical approach*. Cambridge University Press, 2003.

[45] S. Eger, J. Daxenberger, and I. Gurevych. "Neural End-to-End Learning for Computational Argumentation Mining". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by R. Barzilay and M.-Y. Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 11–22.

[46] P. Ekman and W. V. Friesen. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Malor Books, 2003.

[47] L. Gienapp, B. Stein, M. Hagen, and M. Potthast. "Efficient Pairwise Annotation of Argument Quality". 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. 5772–5781.

[48] M. Glockner, Y. Hou, P. Nakov, and I. Gurevych. "Missci: Reconstructing Fallacies in Misrepresented Science". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand, Aug. 2024, pp. 4372–4405.

[49] P. Goffredo, E. Cabrio, S. Villata, S. Haddadan, and J. Torres Sanchez. "DISPUTool 2.0: A Modular Architecture for Multi-Layer Argumentative Analysis of Political Debates". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.13 (July 2024), pp. 16431–16433.

[50] P. Goffredo, M. Chaves, S. Villata, and E. Cabrio. "Argument-based Detection and Classification of Fallacies in Political Debates". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 11101–11112.

[51] P. Goffredo, S. Haddadan, V. Vorakitphan, E. Cabrio, and S. Villata. "Fallacious Argument Classification in Political Debates". In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. Ed. by L. D. Raedt. Main Track. International Joint Conferences on Artificial Intelligence Organization, July 2022, pp. 4143–4149.

[52] 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).

[53] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. Spiecker genannt Döhmann, and C. Burchard. "Mining legal arguments in court decisions". In: *Artificial Intelligence and Law* 32.3 (June 2023), pp. 1–38.

[54] I. Habernal and I. Gurevych. "Argumentation Mining in User-Generated Web Discourse". In: *Computational Linguistics* 43.1 (Apr. 2017), pp. 125–179.

[55] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, and I. Gurevych. "Argotario: Computational Argumentation Meets Serious Games". en. In: *Proceedings of EMNLP 2017 (System Demonstrations)*. ACL, 2017, pp. 7–12. (Visited on 12/23/2021).

[56] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, and I. Gurevych. "Argotario: Computational Argumentation Meets Serious Games". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 7–12.

[57] I. Habernal, P. Pauli, and I. Gurevych. "Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis,

and T. Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[58] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. "Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation". en. In: *Proceedings of NAACL 2018*. ACL, 2018, pp. 386–396.

[59] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. "Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by M. Walker, H. Ji, and A. Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 386–396.

[60] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. "The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

[61] S. Haddadan. "Argument Minining and its application in Political Debates". English. PhD thesis. Unilu - University of Luxembourg, Esch sur alzette, Luxembourg, 2022.

[62] S. Haddadan, E. Cabrio, and S. Villata. "DISPUTool – A tool for the Argumentative Analysis of Political Debates". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 6524–6526.

[63] S. Haddadan, E. Cabrio, and S. Villata. "Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4684–4690.

[64] D. Hample. *Arguing: Exchanging reasons face to face*. Routledge, 2005.

[65] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3309–3326.

[66] P. He, J. Gao, and W. Chen. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. 2021.

[67] P. He, X. Liu, J. Gao, and W. Chen. "Deberta: Decoding-enhanced bert with disentangled attention". In: *arXiv preprint arXiv:2006.03654* (2020).

[68] B. Heinzerling and M. Strube. "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J.

Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[69] C. Helwe, T. Calamai, P.-H. Paris, C. Clavel, and F. Suchanek. *MAFALDA: A Benchmark and Comprehensive Study of Fallacy Detection and Classification*. 2024.

[70] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667.

[71] D. Hovy. "The Social and the Neural Network: How to Make Natural Language Processing about People again". In: *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics. 2018, pp. 42–49.

[72] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021.

[73] Y. Hu, M. Hosseini, E. Skorupa Parolin, J. Osorio, L. Khan, P. Brandt, and V. D'Orazio. "ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 5469–5482.

[74] W. J. Hutchins. "Machine translation: A brief history". In: *Concise history of the language sciences*. Elsevier, 1995, pp. 431–445.

[75] N. Indurkhya and F. J. Damerau. *Handbook of natural language processing*. Chapman and Hall/CRC, 2010.

[76] S. Jackson. "Reason-Giving and the Natural Normativity of Argumentation". In: *Topoi* 38.4 (2019), pp. 631–643.

[77] M. Janier and P. Saint-Dizier. "Argument mining: A survey". In: *Computational Linguistics* 47.1 (2021), pp. 97–135.

[78] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. *Mistral 7B*. 2023.

[79] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. *Mixtral of Experts*. 2024.

[80] J. Jiang, X. Ren, and E. Ferrara. "Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks". In: *Proceedings of the International AAAI Conference on Web and Social Media* 17.1 (June 2023), pp. 459–469.

[81] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, and B. Schölkopf. "Logical fallacy detection". In: *arXiv preprint arXiv:2202.13758* (2022).

[82] K. S. Jones. "Natural language processing: a historical review". In: *Current issues in computational linguistics: in honour of Don Walker* (2001), pp. 2–10.

[83] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, and I. Gurevych. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Ed. by D. Zhao. Santa Fe, New Mexico: Association for Computational Linguistics, Aug. 2018, pp. 5–9.

[84] A. Komninos and S. Manandhar. "Dependency Based Embeddings for Sentence Classification Tasks". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Knight, A. Nenkova, and O. Rambow. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1490–1500.

[85] S. Kraus, J. F. Kennedy, and R. M. Nixon, eds. *The Great Debates: Kennedy vs. Nixon, 1960*. Nachdr. d. Ausg. 1962. Bloomington: Indiana University Press, 1977.

[86] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage, 2004.

[87] D. Kuhn. "Critical thinking as discourse". In: *Human Development* 62.3 (2019), pp. 146–164.

[88] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[89] 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: *CoRR* abs/1909.11942 (2019).

[90] J. Lawrence and C. Reed. "Argument mining: A survey". In: *Computational Linguistics* 45.4 (2020), pp. 765–818.

[91] J. Lawrence, J. Visser, and C. Reed. "An Online Annotation Assistant for Argument Schemes". In: *Proceedings of the 13th Linguistic Annotation Workshop*. Ed. by A. Friedrich, D. Zeyrek, and J. Hoek. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 100–107.

[92] J. Lawrence, J. Visser, and C. Reed. "Harnessing rhetorical figures for argument mining". In: *Argument & Computation* 8.3 (2017), pp. 289–310.

[93] 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 (Sept. 2019), pp. 1234–1240. ISSN: 1367-4803.

[94] S. Lewandowsky, U. Ecker, C. Seifert, N. Schwarz, and J. Cook. "Misinformation and Its Correction Continued Influence and Successful Debiasing". In: *Psychological Science in the Public Interest* 13 (Dec. 2012), pp. 106–131.

[95] M. Lewiński and D. Mohammed, eds. *Argumentation in Political Deliberation*. John Benjamins, 2015.

[96] M. Lewiński and D. Mohammed. "Argumentation in political deliberation." In: *Journal of Argumentation in Context* 2.1 (2013).

[97] M. Lewiński and S. Oswald. "When and how do we deal with straw men? A normative and cognitive pragmatic account". In: *Journal of Pragmatics* 59 (2013), pp. 164–177.

[98] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019.

[99] Y. Li, D. Wang, J. Liang, G. Jiang, Q. He, Y. Xiao, and D. Yang. *Reason from Fallacy: Enhancing Large Language Models' Logical Reasoning through Logical Fallacy Understanding*. 2024.

[100] Y. Li, Q. Zhou, Y. Luo, S. Ma, Y. Li, H.-T. Zheng, X. Hu, and P. S. Yu. *When LLMs Meet Cunning Questions: A Fallacy Understanding Benchmark for Large Language Models*. 2024.

[101] M. Lippi and P. Torroni. "Argumentation mining: State of the art and emerging trends". In: *ACM Transactions on Internet Technology (TOIT)* 16.2 (2016), pp. 1–25.

[102] M. Lippi and P. Torroni. "MARGOT: A web server for argumentation mining". In: *Expert Systems with Applications* 65 (2016), pp. 292–303.

[103] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[104] Y. Liu, X. F. Zhang, D. Wegsman, N. Beauchamp, and L. Wang. "POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Ed. by M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1354–1374.

[105] E. Mancini, F. Ruggeri, S. Colamonaco, A. Zecca, S. Marro, and P. Torroni. "MAMKit: A Comprehensive Multimodal Argument Mining Toolkit". In: *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*. Ed. by Y. Ajjour, R. Bar-Haim, R. El Baff, Z. Liu, and G. Skitalinskaya. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 69–82.

[106] E. Mancini, F. Ruggeri, A. Galassi, and P. Torroni. "Multimodal Argument Mining: A Case Study in Political Debates". In: *Proceedings of the 9th Workshop on Argument Mining*. Ed. by G. Lapesa, J. Schneider, Y. Jo, and S. Saha. Online and in Gyeongju, Republic of Korea: International Conference on Computational Linguistics, Oct. 2022, pp. 158–170.

[107] E. Mancini, F. Ruggeri, and P. Torroni. "Multimodal Fallacy Classification in Political Debates". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Y. Graham and M. Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 170–178.

[108] S. Marro, E. Cabrio, and S. Villata. "Graph Embeddings for Argumentation Quality Assessment". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022, pp. 4154–4164.

[109] T. Mayer, E. Cabrio, and S. Villata. "Evidence Type Classification in Randomized Controlled Trials". In: *Proceedings of the 5th Workshop on Argument Mining*. Ed. by N. Slonim and R. Aharonov. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 29–34.

[110] T. Mayer, E. Cabrio, and S. Villata. "Transformer-based argument mining for healthcare applications". In: *Proceedings of the 24th European Conference on Artificial Intelligenc (ECAI 2020)*. IOS Press, 2020, pp. 2108–2115.

[111] L. Meade. "Fallacies–Warning! Deceptive, Hateful Speech Coming Your Way". In: *Advanced Public Speaking* (2021).

[112] S. Menini, E. Cabrio, S. Tonelli, and S. Villata. "Never retreat, never retract: Argumentation analysis for political speeches". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018.

[113] H. Mercier. "Our pigheaded core: How we became smarter to be influenced by other people". In: *Foundations of Human Interaction*. Oxford University Press, 2013, pp. 373–389.

[114] R. Mestre, R. Milicin, S. E. Middleton, M. Ryan, J. Zhu, and T. J. Norman. "M-Arg: Multimodal Argument Mining Dataset for Political Debates with Audio and Transcripts". In: *Proceedings of the 8th Workshop on Argument Mining*. Ed. by K. Al-Khatib, Y. Hou, and M. Stede. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 78–88.

[115] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013.

[116] C. R. Miller. *The Techniques of Propaganda*. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for Learning. Town Hall, Inc., 1939.

[117] R. Mochales and M.-F. Moens. "Argumentation mining". In: *Artificial intelligence and law* 19 (2011), pp. 1–22.

[118] 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*. 2007, pp. 225–230.

[119] G. Morio and K. Fujita. "End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture". In: *Proceedings of the 5th Workshop on Argument Mining*. Ed. by N. Slonim and R. Aharonov. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 11–21.

[120] N. Naderi and G. Hirst. "Argumentation Mining in Parliamentary Discourse". In: *Principles and Practice of Multi-Agent Systems - International Workshops, Revised Selected Papers*. Ed. by M. Baldoni, C. Baroglio, F. Bex, F. Grasso, N. Green, M. Namazi-Rad, M. Numao, and M. T. Suarez. Vol. 9935. Lecture Notes in Computer Science, Springer. 2015, pp. 16–25.

[121] S. Nair, M. Srinivasan, and S. C. Meylan. "Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge". In: *CoRR* abs/2010.13057 (2020).

[122] H. Nakayama. *seqeval: A Python framework for sequence labeling evaluation*. Software available from https://github.com/chakki-works/seqeval. 2018.

[123] 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.

[124] N. B. Ocampo, E. Cabrio, and S. Villata. "Playing the Part of the Sharp Bully: Generating Adversarial Examples for Implicit Hate Speech Detection". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2758–2772.

[125] R. Paul and L. Elder. *The miniature guide to critical thinking concepts and tools*. Rowman & Littlefield, 2019.

[126] A. Payandeh, D. Pluth, J. Hosier, X. Xiao, and V. K. Gurbani. "How Susceptible Are LLMs to Logical Fallacies?" In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue. Torino, Italia: ELRA and ICCL, May 2024, pp. 8276–8286.

[127] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[128] A. Peldszus and M. Stede. "From argument diagrams to argumentation mining in texts: A survey". In: *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7.1 (2013), pp. 1–31.

[129] J. Pennington, R. Socher, and C. Manning. "GloVe: Global Vectors for Word Representation". 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. 1532–1543.

[130] 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, Volume 1 (Long Papers)*. Ed. by M. Walker, H. Ji, and A. Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237.

[131] S. T. Piantadosi, H. Tily, and E. Gibson. "The communicative function of ambiguity in language". In: *Cognition* 122.3 (2012), pp. 280–291.

[132] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. "Improving language understanding by generative pre-training". In: *OpenAI Report* (2018).

[133] C. Reed, K. Budzynska, R. Duthie, M. Janier, B. Konat, M. Koszowy, and O. Yaskorska. "The Argument Web: an online ecosystem of tools, systems and services for argumentation". In: *Philosophy & Technology* 30.2 (2017), pp. 137–160.

[134] 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.

[135] N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, and I. Gurevych. "Classification and Clustering of Arguments with Contextualized Word Embeddings". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 567–578.

[136] R. Rinott, L. Dankin, C. Alzate Perez, M. M. Khapra, E. Aharoni, and N. Slonim. "Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection". 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. 440–450.

[137] R. Ruiz-Dolz, J. Alemany, S. M. H. Barbera, and A. Garcia-Fornes. "Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation". In: *IEEE Intelligent Systems* 36.6 (Nov. 2021), pp. 62–70.

[138] R. Ruiz-Dolz and J. Lawrence. "Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models". In: *Proceedings of the 10th Workshop on Argument Mining*. Ed. by M. Alshomary, C.-C. Chen, S. Muresan, J. Park, and J. Romberg. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1–10.

[139] S. Sahai, O. Balalau, and R. Horincar. "Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 644–657.

[140] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020.

[141] R. Schaefer and M. Stede. "Argument mining on Twitter: A survey". In: *it-Information Technology* 63.1 (2021), pp. 45–58.

[142] M. Schuster and K. Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[143] 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). Ed. by L. Lee, M. Johnson, B. Roark, and A. Nenkova, pp. 403–419.

[144] Z. Sourati, V. P. P. Venkatesh, D. Deshpande, H. Rawlani, F. Ilievski, H.-Â. Sandlin, and A. Mermoud. *Robust and Explainable Identification of Logical Fallacies in Natural Language Arguments*. 2023.

[145] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, and I. Gurevych. "ArgumenText: Searching for Arguments in Heterogeneous Sources". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Ed. by Y. Liu, T. Paek, and M. Patwardhan. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 21–25.

[146] C. Stab and I. Gurevych. "Annotating argument components and relations in persuasive essays". In: *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*. 2014, pp. 1501–1510.

[147] C. Stab and I. Gurevych. "Parsing Argumentation Structures in Persuasive Essays". In: *Computational Linguistics* 43.3 (Sept. 2017), pp. 619–659.

[148] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. "brat: a Web-based Tool for NLP-Assisted Text Annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by F. Segond. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 102–107.

[149] W. L. Taylor. ""Cloze Procedure": A New Tool for Measuring Readability". In: *Journalism & Mass Communication Quarterly* 30 (1953), pp. 415–433.

[150] G. Team et al. *Gemma: Open Models Based on Gemini Research and Technology*.

[151] 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*. Ed. by P. Koehn and R. Mihalcea. Singapore: Association for Computational Linguistics, Aug. 2009, pp. 1493–1502.

[152] C. W. Tindale. *Fallacies and argument appraisal*. Cambridge University Press, 2007.

[153] C. W. Tindale. *Rhetorical Argumentation: Principles of Theory and Practice*. SAGE Publications, 2004.

[154] 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.

[155] J. Van de Pol, M. Volman, and J. Beishuizen. "Scaffolding in teacher–student interaction: A decade of research". In: *Educational Psychology Review* 22.3 (2010), pp. 271–296.

[156] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H.

Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017, pp. 6000–6010.

[157] P. Vijayaraghavan and S. Vosoughi. "TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3433–3448.

[158] J. Visser, B. Konat, R. Duthie, M. Koszowy, K. Budzynska, and C. Reed. "Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction". In: *Lang. Resour. Evaluation* 54.1 (2020), pp. 123–154.

[159] J. Visser, J. Lawrence, and C. Reed. "Reason-checking fake news". In: *Commun. ACM* 63.11 (2020), pp. 38–40.

[160] V. Vorakitphan, E. Cabrio, and S. Villata. ""Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by R. Mitkov and G. Angelova. Held Online: INCOMA Ltd., Sept. 2021, pp. 1498–1507.

[161] V. Vorakitphan, E. Cabrio, and S. Villata. "Protect: A pipeline for propaganda detection and classification". In: *CLiC-it 2021-Italian Conference on Computational Linguistics*. 2022, pp. 352–358.

[162] H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, and B. Stein. "Building an Argument Search Engine for the Web". In: *Proceedings of the 4th Workshop on Argument Mining*. Ed. by I. Habernal, I. Gurevych, K. Ashley, C. Cardie, N. Green, D. Litman, G. Petasis, C. Reed, N. Slonim, and V. Walker. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 49–59.

[163] 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*. Ed. by M. Lapata, P. Blunsom, and A. Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1117–1127.

[164] D. N. Walton. *Argumentation Schemes for Presumptive Reasoning*. Psychology Press, 1996.

[165] D. Walton. *A Pragmatic Theory of Fallacy*. Studies in rhetoric and communication. University of Alabama Press, 1995.

[166] D. Walton. *Argumentation Schemes*. Cambridge University Press, 2008.

[167] D. Walton. *Informal Fallacies : Towards a Theory of Argument of Criticisms*. Philadelphia: John Benjamins Publishing Company, 1987.

[168] D. Walton. *Media argumentation: Dialectic, persuasion and rhetoric*. Cambridge University Press, 2007.

[169] D. Walton. "Why Fallacies Appear to Be Better Arguments than They Are". In: *Informal Logic* 30 (July 2010).

[170] H. Wang, M. S. Hee, M. R. Awal, K. T. W. Choo, and R. K.-W. Lee. "Evaluating GPT-3 generated explanations for hateful content moderation". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2023, pp. 6255–6263.

[171] A. Weston. *A rulebook for arguments*. Hackett Publishing, 2018.

[172] D. Zarefsky. "Strategic maneuvering in political argumentation". In: *Argumentation* 22 (2008), pp. 317–330.

[173] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "BERTScore: Evaluating Text Generation with BERT". In: *CoRR* (2019).

[174] V. Zurloni and L. Anolli. "Fallacies as Argumentative Devices in Political Debates". In: *Multimodal Communication in Political Speech. Shaping Minds and Social Action*. Ed. by I. Poggi, F. D'Errico, L. Vincze, and A. Vinciarelli. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 245–257.