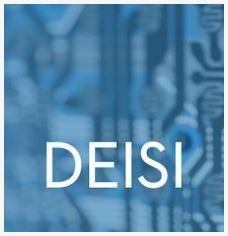

UNIVERSIDADE
LUSÓFONA

A GPT-Based Data Augmentation Approach for enhancing Argumentation Mining in Multi-party Dialogues

Luiza Vidal Copolillo Coelho

Report to obtain the Degree of Bachelor in

Computer Science Engineering

Adviser: Zuil Pirola
Associate Professor, Lusófona University Lisbon
Co-adviser: Manuel Pita
Associate Professor, Lusófona University Lisbon

June, 2024

A GPT-Based Data Augmentation Approach for enhancing Argumentation Mining in Multi-party Dialogues

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my amazing advisors, Prof. Zuil Pirola and Prof. Dr. Manuel Pita, for their invaluable guidance, feedback, and continuous support. I appreciate their expertise and dedication to this work, their belief in me, and their encouragement to always give my best and keep learning.

I would also like to thank all of my closest friends, especially the ones this course has given me, Ana Weng, Joana Okica, and Joana Gonçalves, for their constant encouragement, strength, and dedication to finishing this course. I couldn't have made it without you, and I am so happy that we met and are finishing this chapter of our lives together.

Even from far away in Brazil, I thank my best friend, my heart sister, Nathalia, who is always there for me no matter what. We have grown so much together in these 13 years of friendship, and I wish we could be together to celebrate this victory.

I am also deeply thankful to my girlfriend, Inês, who has helped and supported me in ways that I can't even express. I am so grateful, proud, and fulfilled by you and our love.

I owe a great debt of gratitude to my family. Their support, encouragement, and love have been my sources of motivation and strength. I wouldn't be here to complete this journey if it weren't for my parents, Marcos and Tatiana, and their efforts to support my studies. My grandmothers, whom I admire and miss a lot, have also been great sources of inspiration. I also thank my sisters, Julia and Clara, for their constant support and ability to make everything lighter. And, to my furry baby, Bob, who is my little point of peace when I need it the most.

”

“*Knowing yourself is the beginning of all wisdom.*”

— Aristotle, Philosopher and Scientist

Abstract

This thesis presents a comprehensive study on utilizing Large Language Models (LLMs), specifically GPT-4, for argumentation mining in multi-party dialogues. The primary objective is to investigate whether automated annotation techniques can effectively replace manual annotation processes in the context of argumentation mining. The research employs a zero-shot prompt engineering approach for data annotation to achieve high inter-annotator agreement (IAA) and explores the generation of synthetic corpora annotated by both GPT-4 and human annotators. Through rigorous evaluation, the study reveals that while GPT-4 demonstrates significant potential as an annotator, offering a promising alternative to manual methods, the quality of annotations is notably lower, with a further decline observed in the generated datasets. The findings underscore the advantages of automated annotation in terms of efficiency and scalability, but highlight the need for continued refinement to match human-level quality. This research contributes valuable insights into the practical application of LLMs in natural language processing tasks and sets the stage for future work aimed at enhancing automated annotation systems.

RESUMO

Esta tese apresenta um estudo abrangente sobre a utilização de Modelos de Linguagem de Grande Porte (LLMs), especificamente GPT-4, para a extração de argumentação em diálogos multipartidários. O objetivo principal é investigar se as técnicas de anotação automatizada podem substituir eficazmente os processos de anotação manual no contexto da exploração de argumentação. A investigação emprega uma abordagem de engenharia de prontidão zero para a anotação de dados, a fim de alcançar uma elevada concordância entre anotadores (IAA) e explora a geração de corpora sintéticos anotados tanto pelo GPT-4 como por anotadores humanos. Através de uma avaliação rigorosa, o estudo revela que, embora o GPT-4 demonstre um potencial significativo como anotador, oferecendo uma alternativa promissora aos métodos manuais, a qualidade das anotações é notavelmente inferior, com um declínio adicional observado nos conjuntos de dados gerados. Os resultados sublinham as vantagens da anotação automatizada em termos de eficiência e escalabilidade, mas destacam a necessidade de aperfeiçoamento contínuo para igualar a qualidade a nível humano. Esta investigação contribui com informações valiosas para a aplicação prática de LLMs em tarefas de processamento de linguagem natural e prepara o terreno para trabalhos futuros destinados a melhorar os sistemas de anotação automática.

CONTENTS

List of Figures

List of Tables

1

Introduction

An argument in communication is a set of premises leading to a conclusion in order to persuade or inform (van Eemeren et al., 2014). The identification of arguments is pivotal as it facilitates the analysis of reasoning and the quality of discourse, contributing significantly to the understanding of different perspectives within a conversation. Argumentation is used by humans to convey and defend justifiable viewpoints, to help make informed decisions, to understand complex issues, and to foster productive and constructive discussions (Modgil et al., 2013).

Identifying arguments involves recognizing the structure of a statement, where a premise supports a claim. This process requires understanding the context, the purpose behind the communication, and distinguishing between factual statements, opinions, and persuasive elements. The main feature that distinguishes arguments from other discourse structures is the relation of support between the premises and claims. Here is an example of two statements:

- "We should increase funding for public schools because education is the foundation for a successful society. Studies show that well-funded schools lead to better educational outcomes."
- "I love Italian cuisine more than Mexican cuisine."

The first statement is an argument because it presents a claim (*"increase funding for public schools"*) supported by a premise (*"education is foundational for success"*) and evidence/proof (studies about educational outcomes). The second statement, on the other hand, is a personal preference, and it does not constitute an argument because it does not attempt to persuade or provide reasons for a claim.

1.1 Motivation

Argumentation is regularly used in many fields and plays a vital part in several activities. Many professionals, explicitly or implicitly, use argumentation in the decision-making process (Lawrence and Reed, 2020), in their daily activities: to analyze an issue, to identify advantages and disadvantages, and to make informative decisions. Argumentation provides not only what positions/opinions people adopt, but why they adopt them, providing valuable insights and being applicable in various

domains. Argumentation is studied in a wide range of contexts, from legal and political discourse to everyday conversations and educational settings.

In the *legal* domain, argumentation mining helps analyze juridical decisions (Mochales and Moens, 2011, Teruel et al., 2018), by identifying premises, claims, argument structures, and analyzing support and attack relations among the argument components. This assists judges and legal academics in comparing judgements and understanding case outcomes.

In the *political* domain, argumentation mining is used to detect persuasiveness, fallacies, and coherence in candidates' debates and speeches. Annotated political debates provide insights into how political figures construct arguments to persuade the public and other politicians (Walker et al., 2012, Lippi and Torroni, 2016).

In the *education* domain, argumentation mining can enhance students' learning experience, help teachers evaluate and improve teaching methodologies. Previous studies have proposed an innovative argument learning environment (ALES) that supports the development of critical thinking and reasoning skills in students (Abbas and Sawamura, 2011).

Argumentation identification is not restricted to a single domain or field of study; it includes psychology, philosophy, linguistics, and increasingly, computational tools such as Natural Language Processing (NLP) to handle the complexity and nuance of human language. NLP is the study of how computers and human language interact (Reshamwala et al., 2013). It involves programming computers to process and analyze large amounts of natural language data, aiding in tasks such as text analysis, conversational analysis, and argument identification. Annotated data is commonly used to study and develop argumentation mining models. However, collecting this annotated data is challenging. To address this, techniques from other areas of NLP, such as Data Augmentation (DA), can be leveraged to enhance the performance and applicability of these models.

Data augmentation is one of many regularization techniques with the goal of improving model performance (Mumuni and Mumuni, 2022). These techniques are especially beneficial in low-resource scenarios where annotated data is limited (Feng et al., 2021). The main purpose of DA is to improve the performance and generalization of the machine learning models by adding variation and diversity to the training dataset.

Synthetic Data Generation is a powerful technique for DA that involves using Large Language Models, AI systems designed to understand and generate human-like text by combining extensive training on millions of datasets and powerful computational frameworks (Naveed et al., 2023), to generate new, synthetic labeled data that mimics the structure and content of the original dataset. For instance, the Data Augmentation with a Generation Approach (DAGA) method uses a generation-based approach to create synthetic data for sequence labeling tasks (Ding et al., 2020). The process begins by linearizing labeled sequences and then training a model on these sequences to generate new data. This method has been shown to significantly increase data diversity and improve model performance.

However, a major challenge in data augmentation is the accurate labeling of generated datasets. The performance of machine learning models relies on the quality of these labels, and inaccurate or inconsistent labeling can lead to significant degradation in model performance. A. Mumuni and F. Mumuni (2022) highlight that automated labeling processes, while efficient, often introduce

errors and inconsistencies due to their reliance on predefined rules or heuristics. Hence, ensuring the precision of labels in synthetic data is crucial for maximizing the efficacy of DA in training ML models.

One approach to addressing this challenge is using semi-supervised learning techniques, where a pre-trained model is employed to label the synthetic data. These labels are then reviewed and refined by human annotators to ensure their precision (Liu et al., 2021, Feng et al., 2021). This method combines the efficiency of automated labeling with the precision of human review, resulting in high-quality labeled datasets. Furthermore, techniques such as active learning can be used, where the model identifies uncertain predictions for human review, thereby focusing human effort on the most challenging and informative examples.

Additionally, leveraging advanced models like GPT for data labeling can significantly reduce labeling costs while maintaining high performance. Wang et al. (2021) GPT-3 can label data at a fraction of the cost compared to human annotators, and when combined with human review for low-confidence predictions, it can produce highly accurate labeled datasets. This approach not only reduces costs but also accelerates the labeling process, making it a cost-effective solution for large-scale data annotation tasks.

This work will be based on two of the three methodologies proposed by (Ding et al., 2022), namely Prompt-Guided Unlabeled Data Annotation (PGDA) and Prompt-Guided Training Data Generation (PGDG). PGDA entails the creation of task-specific prompts to direct GPT-3 in annotating unlabeled data. In contrast, PGDG involves developing prompts that guide GPT-3 to autonomously generate a dataset and label it, which can be subsequently employed to train small-scale models. These methodologies rely on the capabilities of unsupervised models like GPT to efficiently and cost-effectively produce labeled datasets.

By carefully managing the labeling process and incorporating data augmentation techniques, the advantages of these methods can be fully exploited, leading to enhanced model performance and generalization in argumentation mining tasks. The use of DA methodologies, such as PGDA and PGDG, are crucial for developing efficient and generalizable NLP models. By increasing the variability and diversity of the training data, these approaches help overcome the limitations of low-resource scenarios and significantly improve the performance of argumentation mining models.

1.2 Problem Identification

Identifying arguments poses significant challenges due to the complexity of natural language, the variability in argument structures, contextual dependencies, and communication ambiguity. The process of argument identification is influenced by the type of communication and its medium (i.e., written text, spoken dialogue, or online forums). A data-driven approach was suggested by Cabrio & Villata (2018) to analyze argumentation mining techniques across various domains. Typically, formal debates have more structured arguments and are, thus, more conducive to argumentation mining compared to casual conversations, where the communication is often more implicit, fragmented, and not always centered around argumentative discourse. Dusmanu et al. (2017) highlighted the complexities of applying argumentation mining to Twitter, as opposed to

other domains, due to the platform's heterogeneous data sources. Moreover, arguments on such platforms pose new challenges, such as differentiating between personal opinions and facts, and detecting the source of disseminating information about such facts for the purpose of verifying their provenance.

The challenges of identifying arguments are amplified in multi-party dialogues due to the dynamic interactions between different speakers and their individual intentions and perspectives. Key issues in multi-party dialogues are referenced by Dignum (2003) and Traum (2003). These issues are critical for any system designed for multi-party dialogue and must be carefully considered, ensuring effective and fair argumentation processes:

1. **System Openness:** Dialogues can be either closed, with a fixed number of participants throughout, or open, where participants can join or leave at any time.

2. **Roles:** Unlike dyadic (face-to-face or one-on-one settings) communications, with just a proponent and an opponent, multi-party dialogues can involve multiple proponents and opponents, each with distinct viewpoints. Some participants might be neutral or act as mediators. In terms of communication, while two-party dialogues have one speaker and one listener at a time, multi-party dialogues can have multiple listeners, and potentially multiple speakers overlapping or interrupting the conversation.

3. **Addressing:** The addressing policy in multi-party dialogues should be clear, either involving public broadcasting to all players or targeted broadcasting to selected participants. In multi-party dialogues, it is hard to keep track of who is talking to whom and which arguments are coming from which people.

4. **Turn taking:** The turn-taking mechanism in multi-party dialogues is more complex than in two-party dialogues. The rules governing who speaks and when can significantly affect the dialogue's outcome, especially in persuasion dialogues.

5. **Termination:** Termination rules differ from two-party dialogues. In multi-party settings, a dialogue might end when a majority is convinced, or all players are, or when no consensus is reached. There's also the possibility of the dialogue continuing indefinitely without a conclusion, necessitating a mechanism to end such situations. Additionally, there should be a way to determine a winner or allow for ties in scenarios where no single player dominates.

1.3 Proposed Solution

The work is aimed at addressing the complexities of identifying argumentative language in multi-party dialogues in group chats and enhancing this task by leveraging Large Language Models (LLMs) capabilities and Prompt Engineering. The focal point is the creation of an annotated corpus of high school student conversations in Portugal, encompassing diverse controversial topics. This corpus is novel for the Portuguese language and is intended to be a resource for argumentation mining (AM) research. To counter the challenge of domain dependency in AM models, the project

proposes using a zero-shot prompt engineering approach to annotate the data and maintain high inter-annotator agreement (IAA). Additionally, we aim to use generative AI to produce a synthetic corpus, which will also be annotated by both GPT and human annotators. The main research question is whether an unsupervised model can effectively substitute a human in the annotation task for argumentation mining.

2

Background and Related Work

This chapter begins with the fundamentals of argumentation, followed by a detailed overview of data annotation methodologies. It then explores several advanced machine learning techniques used in argumentation mining, and concludes with a discussion on data augmentation.

As highlighted in Chapter1, argumentation is studied in a wide range of contexts, from legal and political discourse to everyday conversations and educational settings. Research on AM falls into two distinct paradigms. Firstly, closed domain discourse-level, which focuses on identifying argumentative layout of structured debates or argumentative texts (e.g., student essays). Secondly, information-seeking approaches, which aim to identify standalone argumentative statements from sources that are not inherently argumentative (e.g., everyday conversations between students). This work aligns with the second category, focusing on modelling self-contained arguments from various domains, with little or no contextual background, from multi-party dialogues between students.

2.1 Argumentation Fundamentals

Argumentation is a fundamental part of how we as humans pursue with the world and each other. At its core, argumentation is the process of constructing, expressing, and evaluating an abstract thought in a logical and persuasive manner (Rocha et al., 2016). This goes beyond academic boundaries and affects our daily interactions, decisions, opinions, and beliefs. In every aspect of life, from having discussions about worldwide policies, to even deciding what movie to watch, we are constantly engaged in the practice of presenting and arguing.

Eemeren (2014) summarize a few key concepts come forth as the foundational steps in argumentation:

2.1.1 Claim

As introduced by Stephen Toulmin, a claim is the starting point of any argument, representing the idea the arguer wants others to accept. The notion of a claim goes beyond just stating a fact or an opinion; it often includes a solution to a problem or a statement that deserves attention. For instance, the claim *"Freedom of speech is the most important democratic right"* exemplifies a claim, that expresses a subjective judgment about what is significant, ethical, or desirable.

2.1.2 Premise

In Toulmin's framework of argumentation, a premise acts as a foundational element of any argument, serving as a statement or proposition that the argument assumes to be true. A premise provides the basis/background upon which a claim is supported or attacked. A premise is not just an observation; it typically underpins the principle for the argument, setting the stage for the logical progression towards the claim. For instance, the premise "*The ability to freely express thoughts and opinions is fundamental to the functioning of a democratic society.*" This premise supports the claim mentioned in **2.1.1** by providing a foundational belief that the free expression of ideas is essential for democracy. It sets the groundwork for arguing why freedom of speech should be considered the most critical democratic right, by implying that without this freedom, the essential processes of democracy, like informed decision-making and open debate, cannot function effectively.

2.2 Argumentation Markers

Following the analysis of argumentation and its key concepts, it's valuable to dig into one of the practical ways to identify these structures in discourse: the use of argumentation markers, also known as *"ArgMarks"* or *"shell languages"*. These linguistic cues, such as "because", "due to", "as long as", "but", "although" and "however" are crucial in pointing out the presence of an argument component, such as premises and claims.

Eemeren et al. (2007) have highlighted the significance of these markers for understanding argument structures — an argument scheme offers a precise structure for comprehending how premises are connected to a claim in a discourse. For example, the word *"because"* typically introduces a reason or justification (support), aligning with the concept of positively defending a claim by justifying the involved proposition. Meanwhile, the word *"but"* usually points out a limitation or counterargument (attack), reverberating the idea of negatively defending a claim by refuting the proposition. Hence, these markers are crucial in determining whether a premise supports or refutes an imminent claim.

2.3 Argument Schemes

As mentioned in the previous section, an argument scheme is a model that reveals the principle behind an argument. A particular type of argument is outlined by the logical structure and patterns of reasoning. This idea is not just about identifying the components of an argument, but also involves acknowledging the underlying logic that connects these components, therefore being able to make a deeper evaluation of an argument's effectiveness and the developing persuasive arguments.

The concept of an argument scheme was first introduced by Chaim Perelman and Lucie Olbrechts-Tyteca in 1958, since then it became a cornerstone in the study of argumentation theory. The later work of Perelman and Olbrechts-Tyteca (1969) laid the groundwork for a methodical approach to analyzing arguments, facilitating a more refined understanding of how arguments are organized and function, proving itself valuable in fields such as politics, law and public discourse,

where the efficacy of an argument is crucial. These insights solidify the relevance and advantages of argument schemes in understanding and creating a persuasive discourse.

2.4 Data Annotation

2.4.1 Annotation vs. Data Annotation

Annotation refers to the act of adding notes or comments to a text or other forms of data. As discussed by Marshall (1998), this idea is centered on improving comprehension and offering layers of interpretation to material across diverse domains such as literature, linguistics, and law.

On the other hand, in ML, data annotation is defined as the process of labeling or classifying unprocessed data such as texts, images or even audio (Denny and Spirling, 2018). This process is crucial for the creation of supervised ML models, as it supplies the labeled datasets needed for training and predictions. The precision of these annotations significantly affects the performance of the models, highlighting the pivotal role of data annotation in this area.

2.4.2 Inter-Annotator Agreement

In this work, the data will be manually annotated by different individuals, therefore is important to emphasize the consistency and reliability of the annotation across the different annotators. This concept is known as the Inter-Annotator Agreement (IAA) or Inter-Rater Reliability. Artstein and Poesio (2008) highlight that high IAA indicates clear and objective guidelines and a mutual understanding of these guidelines among annotators. This aspect is particularly crucial in projects involving multiple annotators to ensure the integrity of the data.

Wacholder et al. (2014) later underscore the challenges in achieving high IAA in multi- party dialogues. The work points out the complexities in annotating conversational data, highlighting that nuances and contextual interpretations differ significantly among annotators. This indicates the need for robust agreement metric in such scenarios, to guarantee that the annotated data precisely reflects the many facets of human conversation.

2.4.3 Krippendorff's Alpha Metric

In data annotation, the Krippendorff's alpha stands out as a significant metric for assessing IAA. This metric is particularly recognized for its flexibility in handling numerous data types and measurement levels, including nominal, ordinal, interval, and ratio scales.

Klaus Krippendorff (2011) emphasized statistical robustness and versatility of his methodology in different research scenarios, especially when dealing with missing data and different sample sizes, by providing a detailed explanation of the calculation and application of Krippendorff's alpha (α).

Stab and Gurevych (2014), specifically focus on the use of Krippendorff's alpha in the context of annotating multi-party dialogues. Their work highlights the challenges in such annotation tasks and

underscores the importance of using a robust IAA metric like Krippendorff's Alpha to guarantee reliable and consistent annotations.

2.4.4 Data Annotation Tools

Developed by Pontus Stenetorp, Sampo Pyysalo and Goran Topić, Brat Rapid Annotation Tool (BRAT) is a web-based tool designed for NLP-assisted text annotation, where the notes are structured in a format that allows automatic processing and interpretation by a computer[1].

BRAT is widely recognized in the research community for its intuitive UI, rich features, and adaptability to diverse research needs. It simplifies efficient text data annotation, allowing the identification of spans, relationships, and events, making it especially appropriate for projects that require precise and structured annotations (Stenetorp et al., 2012).

2.4.5 Why Annotate Data?

The difference between the concepts mentioned in section **3.1.1** is key to understanding the importance of data annotation in ML. Whereas traditional annotation boosts a subjective interpretation to a content, data annotation involves assigning objective labels for the purpose of algorithmic comprehension and learning. Hence, the phase *Corpus Annotation* is essential in this work, as it is a foundational stage in developing a supervised ML model.

2.4.6 Challenges of Data Annotation

The Manual data annotation task presents several significant challenges, such as:

- **Labor-Intensive Nature:** This task requires human annotators to manually review and interpret large volumes of data, which can be time-consuming. Each piece of data must be analyzed to ensure that the annotations are precise and accurate, which is essential for the performance of the ML models. Additionally, the complexity of the data can vary, often requiring specialized knowledge or training to annotate correctly. The need for consistency and attention to detail further adds to the workload, as even minor errors in annotation can significantly impact the model's effectiveness.

- **Potential for Occasional Errors:** Human annotators are prone to making errors when dealing with complex data. In addition, distraction and task tedium can also increase the likelihood of mistakes. Even if done unconsciously, personal biases and perspectives can impact judgment in subjective annotation tasks.

- **Difficulty in Training New Annotators:** Training new annotators to consistently and accurately understand the guidelines is challenging and time-consuming, even with detailed instructions.

[1] https://brat.nlplab.org/

- **Data Complexity:** Annotating multi-party dialogues presents unique challenges due to the dynamic and context-dependent nature of conversations between more than two individuals. Different annotators may interpret the same dialogue or message differently, leading to inconsistencies.

2.5 Supervised Model

Supervised learning is a type of ML model trained using a labeled dataset (Sarker, 2021). This technique involves mapping inputs to corresponding outputs and applying this mapping to predict the outputs of unseen data. Hence, the precision of the annotated data used for training significantly affects the performance of the model. It is essential for the model to effectively generalize from the data it was trained on to perform well on unseen data.

The work described by Trautmann et al. (2020), extended models to perform on a more fine-grained level of sequence labeling called Argument Unit Recognition and Classification (AURC). This work studies argument mining considering topic-dependency and accounts for argument stance from heterogeneous sources. Compared to their work, our approach will also explore cross-domain topics but in multi-party dialogues which amplifies the challenges.

2.6 Unsupervised Model

Unsupervised Learning is a type of machine learning model trained using an unlabeled dataset (Sarker, 2021). As opposed to supervised learning, this type of model does not receive the expected results, having to discover possible relationships between the data on its own without human interference. This model is widely used to extract generative features, to discover similarities and differences in data, to identify significant patterns and structures, and for exploratory purposes.

Persing and Ng (2020) introduced an innovative unsupervised method for argument mining, particularly focused on analyzing persuasive student essays. Their approach starts with bootstrapping a small dataset of arguments, using a combination of reliable contextual cues and straightforward heuristics. These heuristics are based on factors like the number of paragraphs, sentence placement, and context n-grams. Once this initial set of labels is established, they train the model in a self-training manner. However, this method predominantly relies on sparse symbolic linguistic features, such as word-based data, which limits its scope to primarily textual elements of the arguments.

The results demonstrated that the unsupervised system outperformed the two supervised baselines (PIPE and ILP), in both the Argument Component Identification (ACI) and Relation Identification (RI) tasks, when using approximate matching. However, their unsupervised system was outperformed by Eger et al.'s supervised system (2017). Overall, the study demonstrates the effectiveness of the unsupervised approach in argument mining, particularly in scenarios where labeled data is not available, although there is still a performance gap compared to more advanced supervised methods like Eger's supervised system.

The study also showed that all models had lower scores for RI tasks compared to ACI tasks. The RI task is inherently more challenging because it requires correctly identifying premises and

claims, and the relationship between them. Experiments demonstrated that removing features specifically designed for the RI task results in only a slight decrease in performance. To improve RI detection there is a need for a deeper semantic understanding of Argument Components (ACs) and the development of intricate boundary detection rules, because argumentative relations often aren't indicated by clear discourse markers, especially when they occur between non-adjacent sentences.

In summary, while the unsupervised approach shows promise, particularly in approximate matching, there are notable challenges in exact AC boundary detection and RI task performance. Future work could benefit from developing a deeper semantic understanding to improve the detection of argumentative relationships and exploring semi-supervised methods trained on annotated data with a wide variety of sentence structures.

2.7 Supervised vs. Unsupervised Comparison

One of the main limitations of using supervised learning for argumentation mining is the labor-intensive and time-consuming process of manually annotating data that is required for training. In multi-party dialogues, this limitation can be aggravated by needing to annotate each segment in the dialogue that may involve more complex, overlapping conversations. This led to the development of unsupervised approaches that do not require labeled data to learn, which reduces the load of producing manual annotations.

Another limitation is that supervised models are usually domain dependent, meaning they are trained for specific domains and struggle to generalize effectively across diverse domains. Some works have introduced cross-domain datasets to mitigate this limitation for supervised models and other works have introduced unsupervised models, such as the work by Persing and Ng described in section 2.6, that are good for identifying hidden patterns or structures that might not be evident in a labeled dataset. However, also described in section 2.6, one of the most serious drawbacks of unsupervised methods is that they focus mainly on lexical cohesiveness, relying solely on words for analysis, and often overlooking non-lexical features which are critical aspects of communication.

Effective analysis of argumentation mining requires considering features beyond words that, usually, unsupervised models do not consider. This includes multimodal cues and multi-party interaction dynamics such as turn-taking and speaker responses. Generally, supervised models can achieve higher accuracy in more complex tasks because these are trained in labeled datasets with a wide variety of sentence structures and, thus, can handle highly nuanced specific argumentative components, that in multi-party dialogues are more complex due to overlapping speeches, interruptions, and subtle cues.

2.8 Large Language Models

Characterized by their deep learning architecture and ability to generate human-like text, Large Language Models (LLMs), represent a significant evolution in natural language processing (NLP) and Artificial Intelligence (AI). ChatGPT, a specialized variation of these models, stands out for its exceptional conversational capabilities and contextual understanding. The development of

ChatGPT involved a two-stage methodology that started with unsupervised pre-training and was then followed by targeted supervised fine-tuning. (Radford et al., 2019).

As highlighted by Hadi et al. (2023) this combination makes it an incredibly effective, versatile, and significant tool in a wide range of applications (e.g. education, legal, healthcare, code writing, finance, and labor market).

Recent studies have outlined the development of AI models, particularly GPT-4 (Achiam et al., 2023). The model marks a notable advancement in NLP and offers significant potential for AM, its key enhancements include:

- **Improved Natural Language Understanding:** GPT-4's sophisticated processing and comprehensiveness of complex language nuances improves the precision in recognizing arguments in casual texts.
- **Scalability and Predictability:** The development of GPT-4 focused on establishing optimization methods that maintain a consistent performance across a variety of settings. This predictable behavior is especially important for the complexities and size variations encountered on AM tasks.
- **Reduced False Information:** GPT-4 has made progress in reducing "hallucinations" (showing a noticeable improvement over the previous GPT-3.5 models) and boosting the reliability of its outputs. This is crucial for mining arguments to ensure that the extracted arguments are accurate.

The enhancements of the GPT-4 model make it ideal for the main task in this work, extracting and understanding arguments from a large dataset.

2.8.1 Prompt Engineering

In the context of LLMs, prompting refers to the process of structuring an input text in a way that effectively guides the AI model to generate desired outputs[2]. Wei et al. (2022) emphasizes the importance of using prompts to obtain advanced reasoning from LLMs. The core idea is to leverage the capabilities inherent in these models, like GPT-4, to solve complex tasks or generate specific responses. This approach contrasts with traditional programming, where explicit logic and data management are coded. Instead, prompting relies on the model's pre-trained understanding of language and concepts to comprehend and respond to the input.

2.8.1.1 Zero-shot Prompting

As mentioned before, prompt engineering is described as the process of crafting inputs to steer the AI model towards more effective responses. A few methods can be used to create a great prompt, this work will primarily employ the *zero-shot prompting* technique among these methods.

[2] https://www.promptingguide.ai/

In zero-shot prompting, the model is presented with a task without any prior examples, the effectiveness of the prompt relies heavily on the clarity of the instruction and the model's pre-existing knowledge. Christiano et al. (2017) suggests that with the integration of human feedback on the model training, LLMs can be fine-tuned to align with human preferences and understanding. This approach potentially enhances zero-shot learning, making AI models more adaptable and responsive to human-like reasoning.

2.8.1.2 Chain of Thought (CoT)

As highlighted before, Wei et al. (2022) introduced the concept of "Chain of Thought" (CoT) prompting, where models are encouraged to generate intermediate reasoning steps when solving complex problems.

The CoT methodology enhances reasoning abilities of language models by guiding them through a logical progression. This approach addresses a common limitation of LLMs, where they might struggle to multitask. By breaking down the task into smaller, manageable steps, CoT helps the model navigate complex problems more effectively.

- **Improved Accuracy:** By decomposing a task into a series of logical steps, the model can tackle each step independently, reducing the likelihood of errors and improving overall accuracy.
- **Enhanced Comprehension:** CoT prompts provide context and clarity, helping the model understand the problem more thoroughly and produce more coherent and relevant responses.
- **Error Reduction:** Intermediate steps allow the model to self-correct and refine its reasoning, minimizing the propagation of errors through the reasoning process.

2.8.1.3 Prompt Elements

- **Task Definition:** defines the core activity or question that AI is expected to address. It sets the stage for the type of problem or query to be discussed.
- **Instructions:** clear and concise directives that guide the AI and its approach to the task. They act as a roadmap of how the AI should process and respond to the prompt.
- **Context:** providing context helps AI understand the nuances and specific circumstances of the task. This can lead the model to better responses.
- **Output Format:** specifies the expected structure of the response.Defining the output format helps in getting the desired form of response.
- **Examples:** providing examples can be beneficial, depending on the nature of the task. They serve as practical demonstrations of the desired form of response, offering AI a clear model to mirror.

2.9 Data Augmentation

The efficiency of machine learning models relies on the quantity, quality, and variety of data. As previously stated, data augmentation is a regularization technique that can be useful in ML and NLP scenarios with limited data (Maharana et al., 2022).

According to Goodfellow et al., 2016, these regularization techniques can avert unwanted model behaviors, such as:

- **Overfitting:** when a model becomes overly familiar with the training data, including outliers, creating varied versions of the training data can help the model generalize better. This exposure to a broader array of scenarios enables the model to perform more effectively on previously unseen data.
- **Lack of Generalization:** introducing variations in the training data (such as new conversational topics) can help the model learn features that are more general and less specific to the training set.
- **Class Imbalance:** if a class is underrepresented in the training data, using DA can balance the dataset, ensuring the model learns to recognize all classes more equally and preventing biased predictions.
- **Insufficient Training Data:** it is quite normal to don't have enough training data available, which can limit the model's performance, by artificially increasing the size of the training dataset, the model gets more data to learn from, improving its robustness and accuracy.
- **Poor performance in real-world scenarios:** models trained on static, clean datasets may perform poorly in dynamic and "noisy" realistic environments, by mimicking real-world variations in the training data, DA helps prepare the model for legitimate applications.
- **Limited variability:** by adding synthetic variations, DA can help the model to handle a broader range of inputs and capture the nuances and complexities of the task.

In argumentation mining, DA holds significant value due to the intricate and diverse nature of argumentative structures, in particular when dealing with multi-party dialogues. Generating diverse argumentative scenarios can enhance the model's ability to overcome the limitations described above across different argumentative structures and topics, this improved reasoning is crucial for developing robust models that can handle real-world argumentation data.

2.9.1 Synthetic Data Generation

Synthetic data generation refers to the process of creating artificial data samples that mimic the real data that is being handled. There are several methods for synthetic data generation, with the use of LLMs, for instance GPT4o, being among the most advanced. These models can generate human-like text, creating (synthetic) data that is both realistic and wide-ranging. The key advantage of using such models is their ability to produce large volumes of data quickly and at a lower cost compared to manual data collection, annotation, and labeling.

2.9.2 Automated Labeling

Automated labeling involves using algorithms and models to annotate data without human intervention. This process can be particularly beneficial when dealing with excessive amounts of unlabeled data. Moreover, LLMs can also be employed for this purpose, giving their understanding of context and language, they can be used to generate accurate labels, therefore, using methods like prompt engineering and employing models like GPT to predict labels based on context and/or examples can significantly reduce time and cost associated with manual labeling, while also improving the consistency and scalability of the labeling process.

2.9.3 Prompt-Guided Unlabeled Data Annotation & Prompt-Guided Training Data Generation

To illustrate the practical application of synthetic data generation and automated labeling, we explore two specific techniques: Prompt-Guided Unlabeled Data Annotation (PGDA) and Prompt-Guided Training Data Generation (PGDG), proposed by Ding et al., 2022.

- **Prompt-Guided Unlabeled Data Annotation (PGDA):** Prompt-Guided Unlabeled Data Annotation (PGDA) is a method where large language models like GPT-3 are used to annotate unlabeled datasets. This technique is valuable in situations where acquiring labeled data is expensive or time-consuming.

  The process of PGDA involves: Designing Annotation Prompts, Data Annotation and Data Training.

  PGDA leverages the pre-training and language understanding capabilities of GPT-3 to produce high-quality annotations. This method is particularly effective for tasks with smaller label spaces and where unlabeled data is available but needs to be labeled.

- **Prompt-Guided Training Data Generation (PGDG):** Prompt-Guided Training Data Generation (PGDG) is an innovative approach where GPT-3 is leveraged to autonomously generate labeled data for specific tasks. This method is particularly beneficial for augmenting data in scenarios where existing labeled datasets are insufficient or imbalanced.

  The key steps in PGDG include: Creations of Prompts, Data Generation Process, Utilization of Generated Data.

2.9.3.1 Comparison: PGDA and PGDG

PGDA is an effective method for annotating existing unlabeled datasets. This approach is particularly beneficial for tasks with a limited number of labels, such as sentiment analysis or simple classification tasks. The primary challenge in these scenarios is not generating new data, but accurately labeling the existing data. This method uses specific prompts to guide LLMs like GPT-3 to annotate data, reducing the dependence on manual labeling.

On the other hand, PGDG focuses on creating new labeled datasets. This technique is especially useful for tasks that involve numerous labels, where a diverse and extensive training dataset is

critical. PGDG involves crafting prompts that instruct GPT-3, or other models, to generate synthetic data that simulates the real data. By expanding the training dataset with synthetic examples, PGDG enhances the model's ability to generalize and perform well on unseen data, addressing issues related to data limitation and improving overall model predictability.

In conclusion, the PGDA and PGDG approaches by Ding et al., 2022 provide efficient and scalable solutions for data augmentation and annotation. By addressing key challenges such as data limitation, class imbalance, and the high cost of manual data annotation, these techniques play a crucial role in the development of more accurate and reliable machine learning models across various NLP applications.

3

SOLUTION

Considering the background and related work reviewed in the previous section, now we present a detailed proposal. The solution has been designed to address the challenges identified in the previous sections and is aligned with the objectives of this scientific study.

This section starts by introducing our approach and expected contributions, followed by a description of the solution's architecture, technologies, and tools used, and scope evaluation.

3.1 Introduction

The objective of this work was to identify argumentative language exchanged in group chats. This task itself is already complicated due to the nuances of natural language. Thus, when we are in a scenario of multi-party dialogues, these complexities increase.

We manually annotated an existing corpus comprising multi-party conversations among high school students from Portugal. This corpus, which is particularly relevant in the context of Portuguese education, included controversial topics such as racism, vaccination, sexism, politics. To the best of our knowledge, no similar corpus exists for Portuguese language. We created an annotated corpus that was thoroughly characterized, and the validation process was carried through an iterative cycle until the IAA was reached. This corpus is expected to serve as a valuable resource for future research and advancements in the realm of AM and DA.

3.1.1 Expected Contributions

- An overview of background work presenting argumentation fundamentals and theory, and related work exploring the different models used for AM.

- To annotate an existing corpus and gain human-level expertise on how to annotate data for AM:

  - Provide a thorough description of the process and understand the annotation iterative cycle until the IAA has been reached.

  - Learn how to impose structure in the source corpus from the annotations and explore the corpus with the new imposed structure by characterizing it statistically and pursuing some qualitative approaches.

- Two different approaches for Data Annotation undergoing the same cross-domain dataset: Manual Annotation and Automated Annotation leveraging LLMs.

- To conduct three separate evaluations, one for the Manual Annotation Module, one for the Automated Annotation Module and another for the Data Augmentation Module. Analyze the results statistically, and compare them to evaluate our proposal.

3.2 Architecture

The solution adopted the following architecture

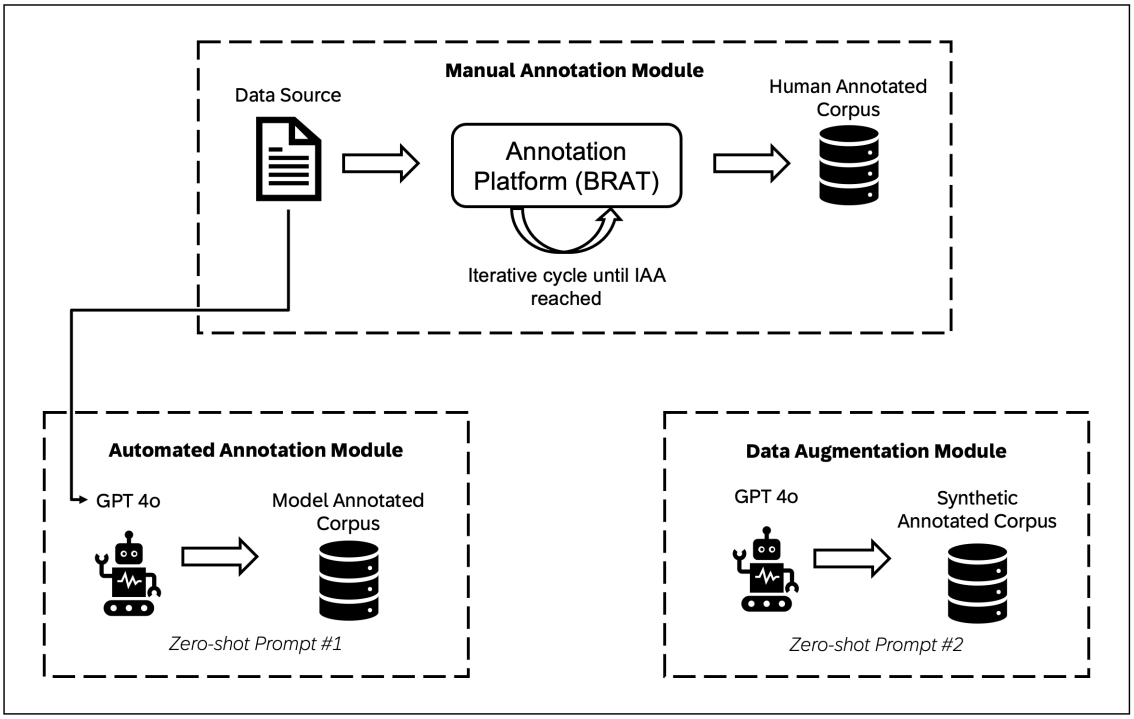

Figure 3.1: Schematic Overview of Proposed Architecture

As represented in Figure 3.1, the architecture of the solution was structured in 3 different data annotation approaches. The process began with the Manual Annotation Module, followed by the Automated Annotation Module. Finally, the sequence culminates with the Data Augmentation Module.

3.2.1 Manual Annotation Module

The Manual Annotation Module was structured into phases that ensured a consistent annotation process. We started with the selection of a discussion room, with a total of 94 turns. We

performed a systematic annotation of this discussion to delve deeper into the data and understand the argumentative discourse we were going to analyze.

Upon the completion of this room, we gathered and manually examined both the annotated and not annotated segments, assessing the IAA between annotators. This analytical session served as a reflection of our annotation practices, allowing us to identify patterns and specificities in the data, and to validate the consistency and accuracy of our annotation schema described in section 3.2.1.1.

Our evaluation led to a satisfying agreement: the foundation we set was both well-planned and significant. After this positive review, we continued to annotate the selected corpus for the Pilot Annotation Phase, including 450 turns in total.

In the Main Annotation phase, we expanded the corpus by doubling the number of rooms, resulting in a combined total of 1769 turns to be analyzed. Given that an evaluation of the annotation practices had already been conducted during the Pilot phase and the guidelines had been refined accordingly, we adhered to these criteria, which once again showed satisfactory results. The guidelines, including the annotation manual and schema in their final version, are detailed in Sections 3.2.1.2 and 3.2.1.1.

After the annotators completed the annotations, we evaluated the results using two methodologies, as described in Sections 5.1.1 and 5.1.2. These sections also detail the metrics employed to interpret, analyze, and discuss the results.

3.2.1.1 Annotation Schema

The annotation schema served as a foundational framework designed to standardize the Annotation Process across different annotators. All the necessary guidelines and schema that the annotators are required to adhere during the process, are encapsulated in an annotation manual, described in the next section.

The annotation tool used to annotate our corpus, is BRAT. Within BRAT, the projects collection configuration has two foundational labels for use:

- **Label "0":** assigned to a turn that is recognized as argumentative but does not have the presence of a premise (i.e., only contains a claim).

- **Label "1":** assigned to a turn that is not only argumentative but also demonstrates a complete structure, containing both claim and premise.

In the Annotation Manual, annotators were instructed to apply these labels by selecting the turn ID (for example, "VAC_R07_000") and then selecting the appropriate label (i.e., 0 or 1). Figure 3.2 exemplifies how the BRAT workspace was organized.

This structure was critical to ensure consistency and reliability in the interpretation and labeling of data, thereby mitigating potential inconsistency in the annotation of data. The approach ensured that the data nature and structure are accurately captured by annotators, which facilitated the subsequent analysis.

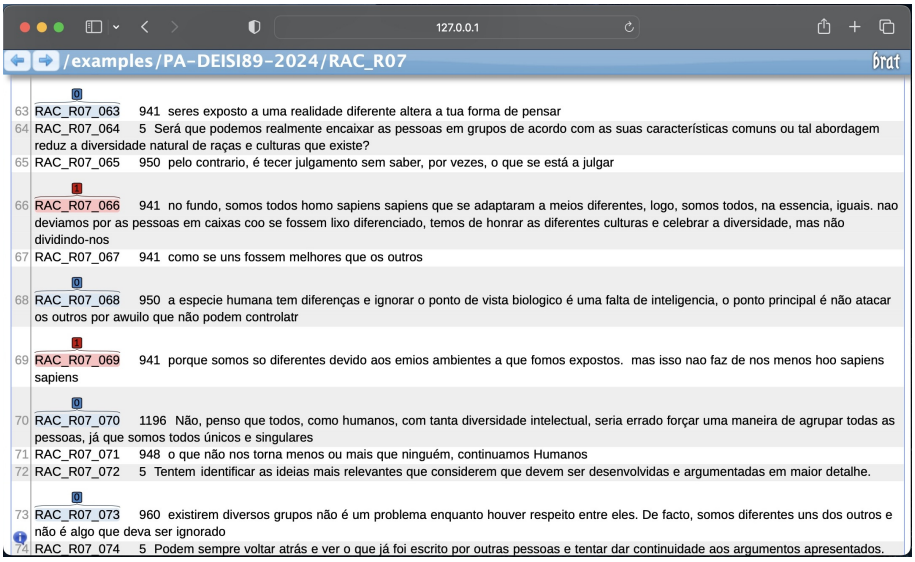

Figure 3.2: Annotator Workspace in BRAT

3.2.1.2 Annotation Manual

The Annotation Manual[1] is a document designed to guide external annotators through the Manual Annotation. It encapsulates all the necessary guidelines and schema which the annotators are required to comply during the execution of the annotations. Thereby, maintaining the consistency and reliability of the annotations. The manual serves several essential functions:

- Task Clarification
- Standardization of the Procedures
- Guidelines for Annotations
- Annotation Schema Descriptive
- Illustrative Examples
- Reference for Inconsistencies

Upon completing the manual and resolving any cases of disagreement, the scope of rooms subject to analysis and annotation was expanded. Following the completion of annotations by both annotators, the classified data was read and processed for thorough evaluation. The methodologies employed, and the results obtained, are detailed in section 5.1.

Manual Annotation Module: Architecture Elements

- **Data Source:** This consisted of selecting the corpus that was later annotated. After annotating a part of the corpus, we set clear criteria and methods for the annotation process. This setup was essential to make sure that we have a solid foundation for our analysis and that all annotators follow the same guidelines.

[1] https://github.com/DEISI-ULHT-TFC-2023-24/TFC-DEISI89-GPTArgMine

- **Annotation Cycle:** This phase was repeated by each annotator until we reached a sufficient IAA to ensure the consistency and reliability of the annotations.
- **Annotated Corpus:** After completing the annotation and calculating the IAA, we will have a human annotated corpus to later analyze, discuss, and compare with our next approach.

3.2.2 Automated Annotation Module

The Automated Annotation Module utilized the same unlabeled data source as the Manual Annotation Module. We used GPT-4o[2], a highly efficient LLM developed by OpenAI, accessing via the OpenAI Platform (Figure 3.3). This model was chosen because of its advanced natural language, context and pattern understanding, and generation capabilities, making it well suited for complex tasks such as argumentation mining.

The first step in this module was pre-processing the data and before sending it as input to the model. This pre-processing step, involved collecting all agreed-upon turns between annotators elaborated during the Manual Annotation module, which was referred to as the *majority* data. Additionally, one of our rules in the manual was not to annotate any turns from the room mediator, to ensure the model wouldn't get confused and make this mistake, we also removed from the dataset any turns where the '*turnID*' parameter was '5'.

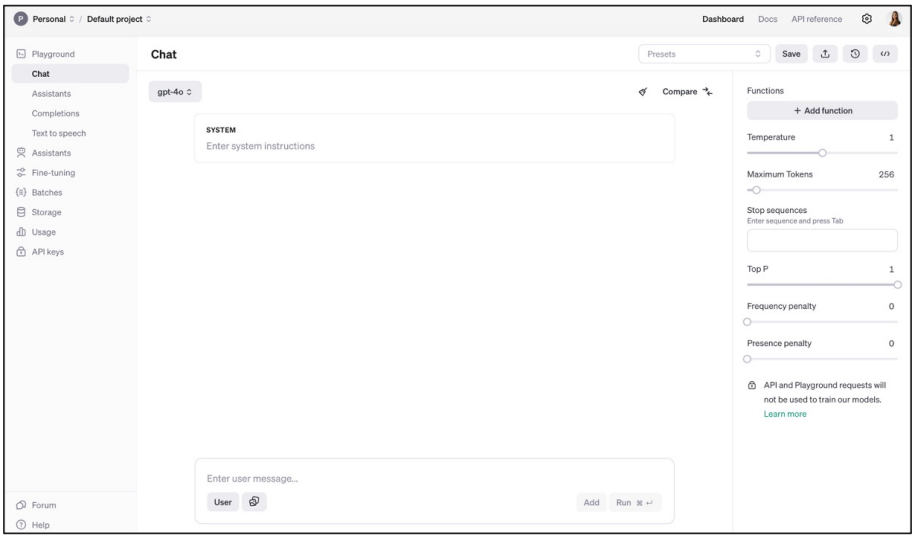

Figure 3.3: OpenAI Workspace

3.2.2.1 Data Balancing

From the majority data, we gathered the total of labels '0' and '1', using Methodology 1 as a reference of classification. To mitigate experimental bias, we balanced the amount of labels, ensuring an equal number of '0s' and '1s'.

Balancing the input data is crucial for accurate and reliable performance metrics in machine learning models. When datasets are imbalanced, the model tends to favor the majority class, leading

[2]https://openai.com/index/hello-gpt-4o/

to misleadingly high accuracy but poor precision, recall, and F1-scores for the minority class. By ensuring balanced data, the model learns to recognize and correctly classify all classes, providing a more accurate and effective evaluation of their performance. Hence, this balance was essential for a better performance of the GPT model across both categories.

3.2.2.2 Prompt Engineering

In our Automated Annotation Module, we applied two techniques of prompt engineering: Zero-shot Prompting. And, as highlighted in Chapter 2, recent research showed that zero-shot prompting, is particularly effective for complex tasks where detailed instructions can leverage the model's inherent reasoning capabilities without overfitting to specific examples.

The final prompt, (Figure 3.4), was crafted to ensure the model followed a logical sequence of reasoning steps. We also specified in the code the desired output format (Figure 3.5).

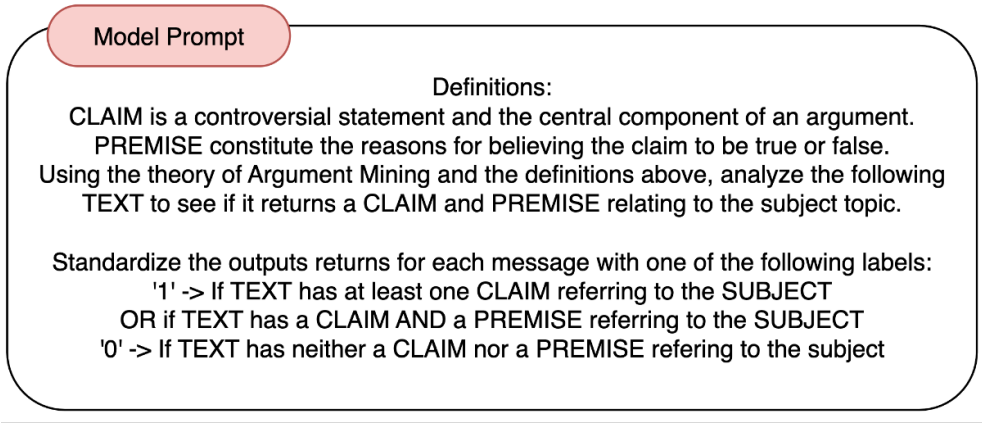

Figure 3.4: Designed Prompt for Automated Annotation Module

By clearly defining the terms claim and premise, we ensured that the model understood exactly what to look for in the text and differentiate between argumentative and non-argumentative content, helping the model's reasoning process. Asking explicitly for the model to standardize the output with specific labels, guaranteed a consistent labeling process, which was vital for the analysis and evaluation.

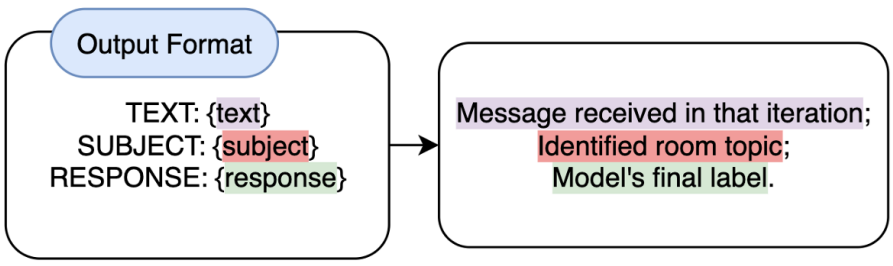

Figure 3.5: Prompt for Automated Annotation Module Output Format

Automated Annotation Module: Architecture Elements

- **Data Source:** We utilized the same data source as the previous module as an input.

- **GPT-4o:** The OpenAI GPT-4o's API was used as the model to perform the automated annotation tasks.
- **Automated Annotated Corpus:** After having the model's annotations, we calculated the IAA, and were left with an annotated corpus that was for analysis and discussion of results.

3.2.3 Data Augmentation Module

The Data Augmentation Module took advantage of Data Augmentation techniques, including synthetic data generation and automated labeling, to artificially expand our dataset. By employing a zero-shot prompting technique alongside the GPT-4o model, we were able to successfully generate a synthetic and automated annotated corpus.

- **Prompt Design:** The prompt was designed to instruct the model to sequentially generate and then classify multi-party dialogues about a controversial theme. It included: background definitions, task instructions, desired nature and format.
- **Generation and Classification:** Upon receiving the prompt, the model produced the synthetic data that adhered to the specifications in the prompt. In addition, the model continued by assigning appropriate labels according to the information it received.
- **Combined Output:** The output from the model included both the synthetic data and the corresponding classifications in a single document. This integrated approach guaranteed that the generated data is immediately useful for evaluation purposes.

Figure 3.6 illustrates the prompt used for this module. We used the same techniques as the previous prompt. However, the model was given more than one instruction: in addition to generating synthetic data for annotation, it was asked to return this data already annotated.

When the annotation was completed, the data was processed according to Methodology 1. This approach allowed us to analyze whether the model could correctly identify turns where a claim was present, without requiring a more in-depth analysis of its ability to differentiate turns with complete argumentative structures (claim + premise).

The manual process of annotating data for 10 chat rooms took approximately 600 minutes. Additionally, the real data collection wasn't entirely under our control, as we depended on external people for this task. In contrast, **generating and annotating** synthetic data for the same number of chat rooms took only 8–10 minutes (approximately). This represents a reduction from 600 minutes to an average of 9 minutes, underscoring the significant reductions in time achieved through automated data augmentation.

The benefits of this approach, besides the enhanced quantity and diversity of data, include *time efficiency* and *scalability*. The automated nature of these processes reduced the time required by approximately 99% compared to manual data collection and labeling. Additionally, the approach allows for continuous data augmentation as new data requirements become evident.

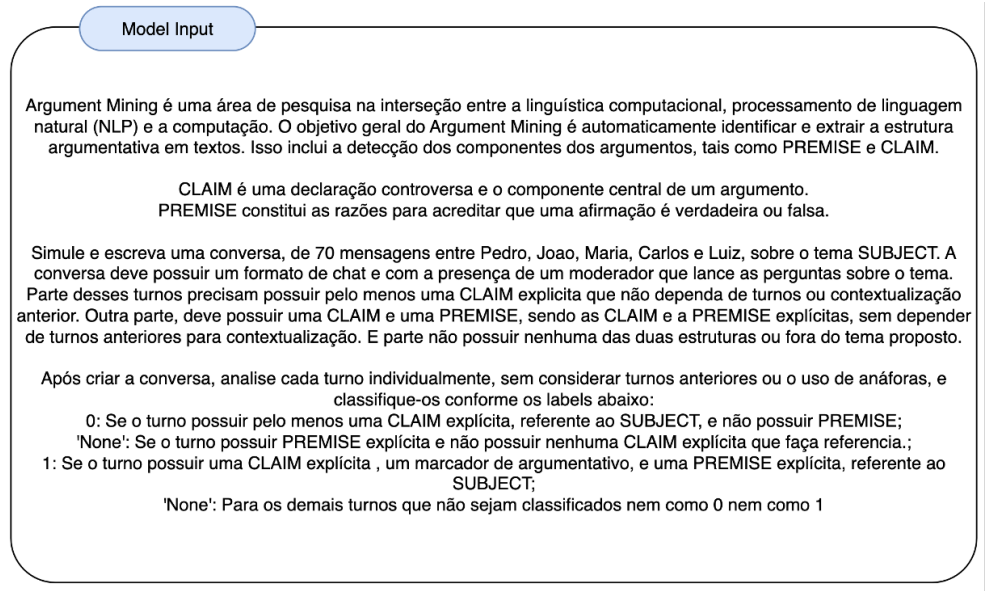

Figure 3.6: Designed Prompt for Data Augmentation Annotation Module

Data Augmentation Module: Architecture Elements

- **GPT-4o:** As the previous module, GPT-4o's API was also used. But this time, the model was asked to generate synthetic data and also perform the annotation.
- **Synthetic Annotated Corpus:** This corpus contained both the model's annotations and human annotations of the previous generated data.

3.3 Technologies and Tools

We chose *Python* as the primary programming language, due to its large ecosystem and range of libraries, which are particularly effective for studies involving data and ML.

The machine learning approach was powered by OpenAI's GPT-4o, a model that has demonstrated state-of-the-art performance in various NLP tasks, leveraging its advanced capabilities with zero-shot prompting. It was selected for its proven effectiveness and compatibility with the requirements and architecture of this project.

Figure 3.7 details the annotation workflow starting with a JSON formatted data source, that will be annotated via BRAT, resulting in an annotated corpus when the IAA is reached, utilizing Krippendorff's Alpha Metric for reliability measurement.

Figure 3.8 illustrates the structure of the Automated Annotation and Data Augmentation Modules.

3.4 Scope

The scope of this thesis is an interdisciplinary venture that integrates concepts from key academic modules and validates their usability in real-world scenarios.

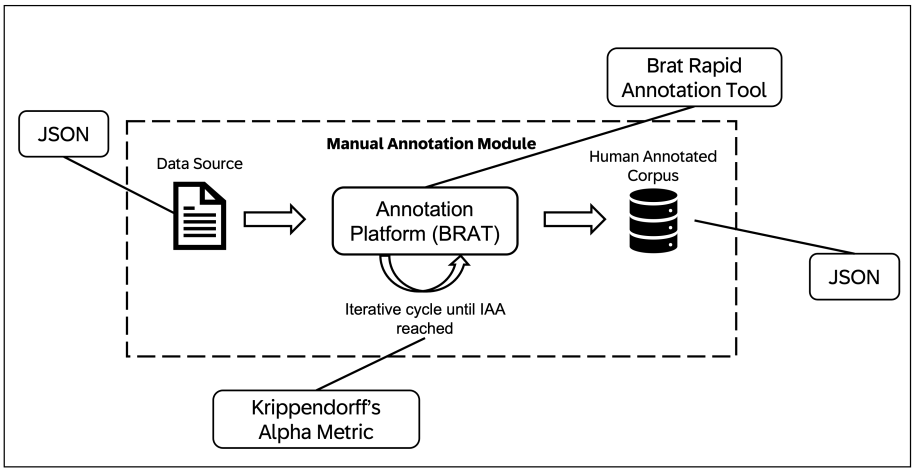

Figure 3.7: Manual Annotation Module Technologies and Tools Breakdown

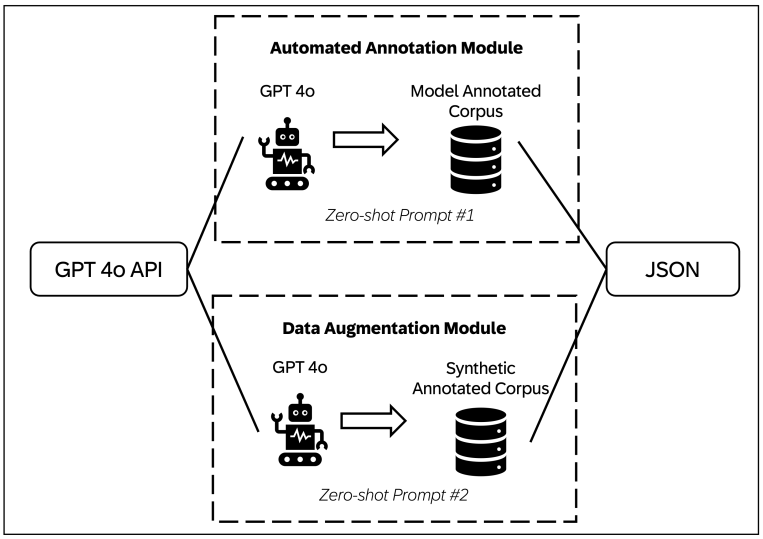

Figure 3.8: Automated Annotation and Data Augmentation Module Technologies and Tools Breakdown

- **Programming Fundamentals:** Scripting data processing, ML, and analyzing tasks within the project.
- **Artificial Intelligence:** Understand the landscape of possible ML approaches. Implementation based on supervised learning, ChatGPT prompting and LLMs.
- **Mathematics I and II:** Explore the corpus with the new imposed structure by characterizing it statistically and pursuing some qualitative approaches.
- **Data Science:** Gain human-level expertise on how to annotate and impose structure on an existing large corpus for ML.
- **Requirements and Testing Engineering:** Defining project specifications and tasks to be performed.

- **Algorithms and Data Structures:** handling and processing the data to ensure robustness and optimized performance.

4

Requirements Gathering and Analysis

Given that this TFC is scientific in nature, the Requirements Gathering and Analysis are approached in a generalized way for the study of the data. The following table lists the project's requirements.

Table 4.1: Requirements Table

ID	Module	Description	Type
REQ1	Manual Annotation Module	Manual Annotation and IAA measuring.	Mandatory
REQ2	Automated Annotation Module	GPT-4o Real Data Annotation and IAA measuring.	Mandatory
REQ3	Data Augmentation Module	GPT-4o Synthetic Data Generation and Labeling.	Mandatory
REQ4	Data Augmentation Module	Manual Annotation o f Synthetic Data and IAA measuring.	Mandatory
REQ5	Data Augmentation Module	Apply Chain of Thought Prompting Technique in Synthetic Data.	Optional

Our work focuses on leveraging Large Language Models and Data Augmentation to identify argumentative language from various domains in multi-party dialogues between students. The corpus for this work was collected from discussions, composed of text in Portuguese, on various topics that occurred in classrooms during class time. In each class, there was a mediator that started, mediated, and ended the discussion, where participants would take turns debating on a specific topic. Here's an example of the start and end of a discussion made by the mediator.

Start of the discussion by Room Mediator

- *(Original - Portguese)* "Olá! Sou o moderador. Neste momento os intervenientes estão a entrar na plataforma. O debate terá início em 1 minutos. Violência de género: a forma de vestir incita ou justifica agressão, sim ou não? Recomendamos escrever frases completas e manter uma atitude de atenção e resposta às intervenções dos outros. Tenta ver os dois lados da discussão, o teu e o de quem não pensa igual a ti. Nesta primeira fase do debate é esperado que os intervenientes contribuam com as suas próprias ideias, assim como estender ou questionar as ideias dos outros. Notem que existe uma opção em forma de seta curva

para responder a alguma mensagem em específico, como no WhatsApp. Recomendamos que usem está opção para interagir diretamente com as intervenções dos vossos colegas."

- *(Translated - English)* "Hello, I'm the moderator. At this moment the participants are entering the platform. The debate will start in 1 minute. Gender violence: does the way we dress incite or justify aggression, yes or no? We recommend writing complete sentences and keeping an attentive and responsive attitude to other people's interventions. Try to see both sides of the argument, yours and those who don't think like you. In this first phase of the debate, participants are expected to contribute their own ideas, as well as extend or question the ideas of others. Note that there is an option in the shape of a curved arrow to reply to a specific message, just like on WhatsApp. We recommend that you use this option to interact directly with your colleagues' interventions."

End of the discussion by Room Mediator

- *(Original - Portguese)* "Vamos encerrar..."

- *(Translated - English)* "Let's finish..."

As mentioned in section 1.2, there are key issues related to argumentation mining in the context of multi-party dialogues: system openness, roles, addressing, turn taking and termination. These issues are critical for any system designed for multi-party dialogue and must be carefully considered. We will describe next how these issues were solved or mitigated:

- **System openness:** The dialogues collected are open, in the sense that participants could enter/leave as they wished, but would not happen frequently because discussions were conducted during class time to mitigate the number of participants leaving/entering dialogues in the middle.

- **Roles:** Each discussion had a mediator who did not advocate any position, but rather started and ended the dialogue, and guided the dialogue facilitating communication between participants.

- **Addressing:** It was recommended in each dialogue to interact and directly address participants and even replying directly to specific messages by using the "curved arrow".

- **Turn Taking:** Each dialogue was labelled and separated by turns where each participant would take turns to intervene. Labelling the dialogue on turn-taking is important because in multi-party dialogues the rules of who speaks and when can significantly affect the dialogue's outcome.

- **Termination:** The mediator would give a 2-minute warning before ending the discussion to give time for the participants to say their final thoughts

The annotation will be at a *turn-level* (i.e. we will be identifying turns that contain argumentative language as a whole, instead of individually identifying the components that comprise an argument – claim + premise) with the aim of increasing the IAA. Consider the following examples of a non-argumentative turn and an argumentative turn.

Argumentative

- *(Original - Portguese)* "Não existem argumentos para justificar isso, uma vez que a roupa que se usa não tem nada a ver com ser agredida sexualmente. Quanto ao facto de as mulheres se taparem, isso já depende dos princípios de cada cultura."
- *(Translated - English)* "There are no arguments to justify this, since the clothes you wear have nothing to do with being sexually assaulted. As for whether women cover up, that depends on the principles of each culture."

Non-Argumentative

- *(Original - Portguese)* "E mesmo que tivessem algum tipo de problema isso não justifica essa ação."
- *(Translated - English)* "And even if they had some kind of problem, that doesn't justify this action."

5

EVALUATION

In this Chapter we will provide in-depth analysis on the results obtained from the three evaluation we performed on each module described in Chapter 3. First, we begin with the results from our Manual Annotation Module as described in Section 5.1, followed by a comparison of the performance of our Manual Annotation Module versus our Automated Annotation Module as described in Section 5.2, and finally a comparison of both modules in annotating the generated data created by our Data Augmentation Module as described in Section 5.3.

5.1 Manual Annotation Module Results

In this section, we explore the two methodologies used to analyze the accuracy and consistency of our Manual Annotation Module described in Section 3.2.1.

Described in section 5.1.1, Methodology 1 employs a binary labeling system to determine if dialogue turns are annotated or not, achieving a high inter-annotator agreement (IAA) score of 0.87.

Furthermore, section 5.1.2 introduces Methodology 2, a more detailed labeling approach to distinguish between complete and incomplete argumentative structures, resulting in a slightly lower IAA score of 0.81. This refined approach highlights the robustness of our annotations under a more detailed examination. Both methodologies contribute significantly to our understanding of the effectiveness and reliability of the annotation guidelines used.

Table 5.1 summarizes the classification of the annotations produced in BRAT.

	Not Annotated	**Annotated as 0**	**Annotated as 1**	**Total**
Annot 1	1347	245	177	1769
Annot 2	1359	240	170	1769
Total	2706	485	347	3538

Table 5.1: Summary of Main Annotation

5.1.1 Methodology 1

In the first part of our analytical process, we set up a binary labeling system to organize the annotated data in a format that would ease a preliminary evaluation.

Each turn within the dialogue was assigned a label reflective of its annotation status:

- Turns that had not been annotated were labeled with '0'.
- Annotated turns, regardless of whether they contain a complete argumentative structure featuring both a 'claim' and a 'premise' (label 1 in BRAT) or include only a 'claim' (label 0 in BRAT), were consistently labeled with '1'.

This analytical methodology was designed to compare our agreement on whether to annotate or not, independent of the argumentative structure (i.e., complete or incomplete argument).

5.1.1.1 Data Analysis and Findings

The culmination of the methodology was evidenced by achieving an **Inter-Annotator Agreement (IAA) score α = 0.87**. This level of agreement underscores the reliability of our annotations and validates the annotation guidelines employed in our initial analysis.

The confusion matrix[1] allowed us to assess the accuracy of our annotations and detect any consistent biases or errors in our labeling and annotation methodology.

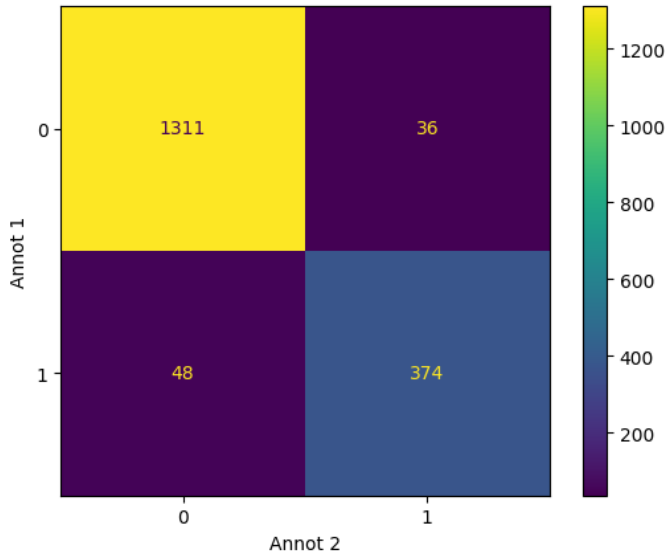

Figure 5.1: Confusion Matrix for Methodology 1

Figure 5.1 presents a confusion matrix that compares the results of the two annotators. Annotator 1 identified 422 turns (374 + 48) as containing argumentative language and did not annotate 1347 turns (1311 + 36). Annotator 2 marked 410 turns (374 + 36) as having argumentative language and

[1] A confusion matrix is a tool used in supervised learning to evaluate the performance of classification models.(Access: Understanding Confusion Matrix)

did not mark 1359 turns (1311 + 48). Together, the matrix reflects the analysis of 1769 instances in total per Annotator (422 + 1347 annotated by Annotator 1 or 410 + 1359 by Annotator 2).

In the confusion matrix it is also visible the number of turns that had an agreement between annotators, with a total of 1685 (1311 + 374) turns, and disagreement between annotators, with a total of 84 (48 + 36) turns.

During our pilot annotation phase, we performed an in-depth analysis of each **disagreement** case individually and made some inferences:

- **Relevance to the Room Topic:** Turns where the statement is not directly related to the topic of discussion caused confusion among annotators. An example is this turn from a room about racism: *(Original - Portguese)* “julgas porque te é intrinseco" *(Translated - English)* “you judge because it’s intrinsic to you". From this sentence, the annotator cannot directly relate it to the topic racism without a given context, thus, the turn should not be classified as an argument.

- **Spelling errors:** Turns where there are spelling errors might lead to misunderstandings. An example is the turn *(Original - Portguese)* “A educação tem um papel preponderante na **formaç**", *(Translated - English)* “Education plays a key role in **format"**, where one annotator understood that the word “formaç" meant *(Original - Portguese)* “formação" *(Translated - English)* “formation", and the other annotator did not understand it. It was agreed afterward that turns that contain spelling errors, but are still comprehensible, should be annotated.

- **Figures of Speech and Language Vices:** Expressions that include figures of speech or language vices can be difficult to interpret and categorize into specific entity labels, thus, should not be annotated. An example is the turn *(Original - Portguese)* “e ideia de **“outro"** tem uma conotação negativa associada a si.", *(Translated - English)* “and the idea of **“other"** has a negative connotation associated with it."

- **Anaphora:** Anaphoras difficult the determination of whether a claim is being made. An example is the following turn *(Original - Portguese)* “**Esse ponto de vista** não é válido já que se todos pensarmos dessa maneira, ninguém se vacina", *(Translated - English)* “**That point of view** is not valid because if we all think that way, nobody gets vaccinated". This turn contains an anaphora ("Esse ponto de vista") that depends on previous mentioned turns, thus, should not be annotated.

- **Turns Containing Only Premises:** Turns that contained only premises and had argument markers starting the sentence (e.g., Porque) caused confusion between annotators. An example is the turn *(Original - Portguese)* “**Porque** a educação varia de pais para pais", *(Translated - English)* ‘**Because** education varies from country to country". This turn contains only a premise and not a claim, hence, should not be annotated.

5.1.2 Methodology 2

Building upon the initial methodology, our subsequent analytical strategy introduced a more granular labeling system to further refine our data evaluation process. In this methodology, every turn in the dialogue was systematically categorized based on its annotation content:

- Turns that had not been annotated were labeled with '0'.
- The turns annotated as argumentative, containing a 'claim' but lacking a 'premise' (label 0 in BRAT), were labeled with '1'.
- Turns annotated as argumentative, containing a complete argumentative structure, 'claim' and 'premise' (label 1 in BRAT), were labeled with '2'.

This labeling schema was selected to sharpen our analysis by differentiating the depth of argumentation within the annotated turns and enabling a more nuanced assessment of our annotation precision.

5.1.2.1 Data Analysis and Findings

The implementation of Methodology 2 led to an **IAA score of α = 0.81**. This IAA score slightly lower than our initial methodology, which is normal, since this analysis is more in-depth. However, this score still represents a strong agreement between annotators, reinforcing the reliability of our annotations and guidelines under a more complex classification system.

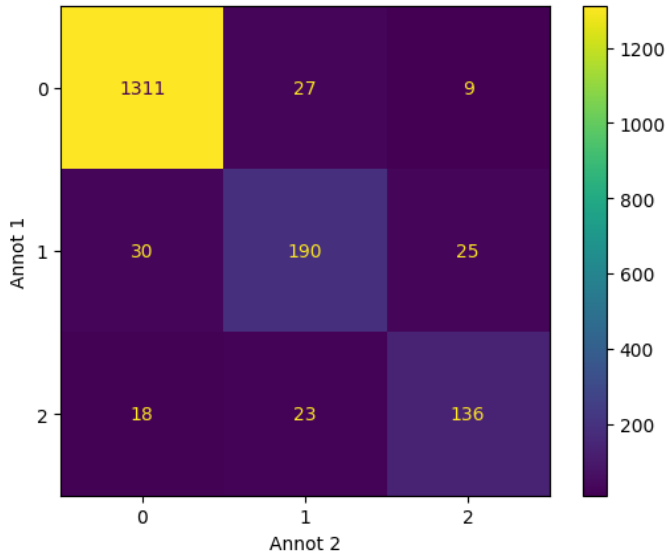

Figure 5.2: Confusion Matrix for Methodology 2

Figure 5.2 displays a comparative analysis of the two annotators' results on Methodology 2. Annotator 1 identified 177 turns (18 + 23 + 136) as containing a fully developed argumentative structure (label '2'), labeled 245 turns (30 + 190 + 25) as an incomplete argumentative turn (label '1'), and did not annotate (label '0') a total of 1347 turns (1311 + 27 + 9). On the other hand,

Annotator 2 identified 170 turns (136 + 25 + 9) as argumentative with a complete argumentative structure (label '2'), 240 (23 + 190 + 27) as argumentative, but lacking the presence of a premise (label '1'), and did not annotate (label '0') 1359 turns (18 + 30 + 1311). Combining these annotations, the matrix encapsulates a total analysis of 1769 instances (177 + 245 + 1347 annotated by Annotator 1 or 170 + 240 + 1359 annotated by Annotator 2).

In the confusion matrix, the number of turns that had an agreement between annotators is also visible, with a total of 1637 (1311 + 190 + 136) turns, and disagreement between annotators, with a total of 132 (27 + 9 + 30 + 25 + 18 + 23) turns.

During our pilot annotation phase, we performed an in-depth analysis of each **disagreement** case individually and noticed the same inferences as described in Subsection 5.1.1.1 and other inferences:

- **Non-conventional vs. Common ArgMarks:** Turns that use non-conventional markers to denote arguments, can make it harder to identify arguments. An example is the turn *(Original - Portguese)* "por vezes nao. ha paises que formam esteriotipos de outrso (nacionalidades), **nao querendo isso dizer que** esse esteriotio seja verdadeiro", *(Translated - English)* "sometimes not. there are countries that form stereotypes of others (nationalities), **not to say that** this stereotype is true".

- **Order of the Argument Components:** The sequence of claim and premise (or vice versa) can be confusing for annotators and make it harder to classify. An example is the turn *(Original - Portguese)* "sim, ate que cesce e se expoe as noticias, a internet, le livros e educa-se de forma diferente. Uma criança fruto de uma familia racista pode nao o ser", *(Translated - English)* "Yes, until you grow up and expose yourself to the news, the internet, read books and educate yourself differently. A child from a racist family may not be one" where the first sentence is the premise and the latter the claim. Nevertheless, these turns should be annotated regardless of the order of components.

5.1.3 Conclusion

Based on the results and analysis described previously, we can conclude that both Methodologies had high inter-annotation agreements **IAA average score of α = 0.84**, showcasing the reliability and effectiveness of the annotation guidelines used by the annotators. It is important to highlight the disagreement cases between annotators and when these occur to establish a higher agreement between annotators for future annotations. The disagreement cases were identified during the Pilot Annotation phase and explicitly described in the Annotation manual (Section 3.2.1.2) to be used as annotation guidelines for all annotators during the Main Annotation phase and for future annotations.

It is also important to mention that **disagreements** between annotators were also due to **confusing turns** (e.g., *(Original - Portguese)* "isso nao quer dizer que um argentino seja a minha ideia de um argentino.... nao sei se me fiz entender", *(Translated - English)* "that doesn't mean that an argentine is my idea of an argentine.... I don't know if I made myself clear") which make it

confusing for annotators to classify. Another disagreement case is **human-prone errors**, due to the exhaustive manual labor of annotation that can lead to errors by mistake.

5.2 Automated Annotation Module

As previously mentioned, one major drawback of a Manual Annotation approach is the time-consuming process of annotating. Employing new automated approaches such as leveraging LLMs, can potentially provide more consistent and reliable annotations while significantly reducing the required manual effort. In this section, we will evaluate the accuracy and consistency of our Automated Annotation Module, compared to the results of our Manual Annotation Module described in Section 5.1.

As described in Section 3.2.2, the Automated Annotation Module started with the **pre-processing of the data**: (1) Collecting all agreed-upon turns between all annotators from the Manual Annotation Module, (2) Gathering the total of labels '0's and '1's using Methodology 1 as described in Subsection 5.1.1, (3) Balancing out the amount of labels (randomly sampling equal amount of '0's and '1's) to mitigate experimental bias, (4) Testing different prompts until reaching the ideal one.

5.2.1 Data Analysis and Findings

We evaluated the GPT-4 model on unlabeled data using a zero-shot prompt approach to assess the accuracy and consistency of the generated annotations. This evaluation was done by employing the methodology 1 described in Subsection 5.1.1 which achieved an **Inter-Annotator Agreement (IAA) score α = 0.72** between the human annotators and the model. This score empirically showcases overall a high degree of agreement between GPT and human annotators, highlighting the potential of LLM models to provide consistent argument quality annotations.

Table 5.2, highlights the comparison of the results between the Manual Annotation Module and the Automated Annotation Module. The IAA score of Automated Annotation is slightly lower than the results of our **Manual Annotation Module (IAA score α = 0.87**). This is expected because generative models like GPT-4 focus mainly on lexical cohesiveness, relying solely on words for analysis, and often overlooking non-lexical features which are critical aspects of communication. Multi-party dialogues between students are more complex to annotate because these contain informal language without a clear CLAIM + PREMISE structure, and also include multimodal cues and multi-party interaction dynamics such as turn-taking, overlapping speeches, interruptions, and subtle cues.

	Time (Min)	**IAA**
Manual Annot	600	0.87
Automated Annot	6	0.72

Table 5.2: Comparison of Manual Annotation and Automated Annotation. Time for manual annotation excludes the time spent on instruction preparation and training.

Based on the table, there is a clear **trade-off between annotation quality and time spent** on annotation. In the manual annotation, each annotator spent approximately 1 hour to annotate one classroom, whereas GPT-4 spent approximately 6 minutes to annotate all classrooms. On the other hand, the IAA score was higher in the manual annotation approach, compared to the automated one. Furthermore, we also evaluated the **F1 score = 0.87**, **Precision score = 0.84**, **Recall score = 0.89** and **Accuracy score = 0.86**.

Figure 5.3 displays a further comparative in-depth analysis of the results of the Manual Annotation Module and the Automated Annotation Module.

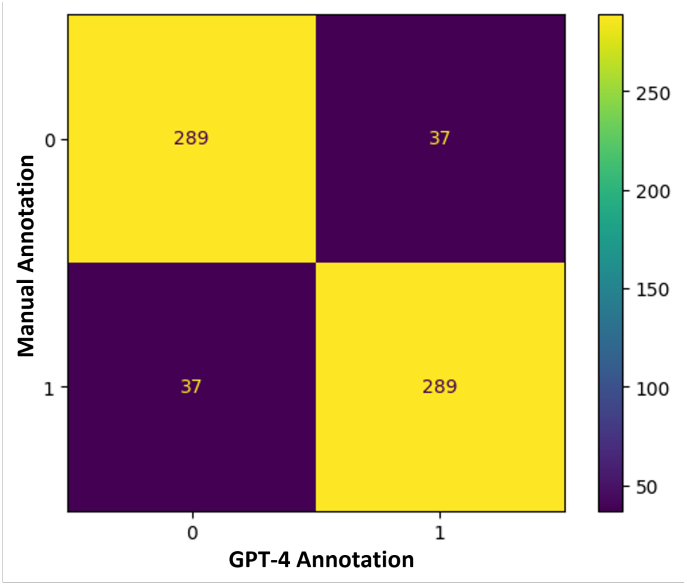

Figure 5.3: Confusion Matrix Manual Annotation vs. Automated (GPT-4) Annotation.

Manual Annotators identified 326 turns (35 + 291) as containing argumentative language and did not annotate 326 turns (271 + 55). GPT-4 marked 346 turns (291 + 55) as having argumentative language and did not mark 306 turns (35 + 271). Together, the matrix reflects the analysis of 652 instances in total per Annotator (326 + 326 annotated by the Manual Annotators or 346 + 306 by GPT-4).

5.2.1.1 Disagreements and GPT's Limitations Discussion

In the confusion matrix, it is also visible the number of turns that had an agreement between the annotators and the model, with a total of 562 (271 + 291) turns, and disagreements, with a total of 90 (35 + 55) turns. We performed an in-depth analysis of the **disagreement** cases and made some inferences:

Contextual Understanding: GPT can sometimes miss implicit context coming from previous interactions in the chat room. Argumentation mining requires considering features beyond words, which GPT does not consider. For instance, the following turns were classified as non-arguments (label '0') by GPT and classified as arguments (label '1') by human annotators: (1) *(Original - Portuguese)* "Estereótipo são opiniões e ideias generalizadas, utilizadas pelas pessoas para

pré-definir alguém ou algo quanto ao seu comportamento, gênero, aparência, religião, cultura, condição social, etc.", *(Translated - English)* “Stereotypes are generalized opinions and ideas used by people to predefine someone or something in terms of their behaviour, gender, appearance, religion, culture, social status, etc."; (2) *(Original - Portuguese)* “cabe a cada um respeitar ou ter preconceito", *(Translated - English)* “it’s up to each person to respect or have prejudice".

Human annotators considered these two turns as viewpoints from the students answering the question asked previously by the room mediator *(Original - Portuguese)* “Os estereótipos e preconceitos fazem parte da natureza humana ou são típicos de certas pessoas ou culturas? Sim ou não? Porquê?", *(Translated - English)* “Are stereotypes and prejudices part of human nature, or are they typical of certain people or cultures? Yes or no? Why?". On the other hand, GPT misses the context from previous interactions in the room and saw these turns as merely definitions without explicit argumentative support. Furthermore, GPT-4 might have stricter criteria for what constitutes a full argument, requiring more explicit premises or detailed support.

Turns Containing Only Premises: As described in 5.1.1.1, one disagreement case found and discussed among human annotators during the Pilot Annotation phase was turns that only contained premises and not claims. There were multiple disagreement cases between GPT and human annotators, where GPT would annotate as ’1’ a turn that contains only a premise and not a claim, whereas annotators would annotate as ’0’. For instance, the turn *(Original - Portuguese)* “**porque** a nossa geracao sozinha nao e capaz de alterar um perconceito tao antigo como o racismo", *(Translated - English)* “**because** our generation alone is not capable of changing a prejudice as old as racism", was annotated as ’0’ by human annotators and ’1’ by GPT. GPT analyzed this turn as a CLAIM about the generational impact of racism, whereas human annotators saw it as a PREMISE starting with an argument marker without a clear argumentative intent. As mentioned previously, this case as been identified in the Manual Annotation module during the Pilot Annotation phase and explicitly described in the Annotation manual (Section 3.2.1.2).

To mitigate these and other cases described in the manual, a hypothesis we formulated was if providing the manual as a basis and asking GPT to annotate the data would improve its accuracy. We tried this approach, we provided GPT with the full annotation manual and the unlabeled data, and asked GPT to return the annotated data. We discovered that GPT performed worse when provided the annotation manual and the unlabeled data simultaneously because it is not designed to multitask effectively. Handling both tasks at the same time could overwhelm the model, leading to reduced performance on both fronts. When the annotation manual is not provided, GPT can focus solely on the primary task of understanding and processing the unlabeled data, which improves its overall effectiveness. For future research, we can evaluate if providing a chain of thought with step-by-step instructions of the annotation manual could potentially help improve GPT’S performance by making the process clearer and easier to follow.

5.2.2 Conclusion

Based on the results and analysis described previously, the **Automated Annotation Module**, which leverages LLMs such as GPT-4, **provides a promising alternative** to the Manual Annotation

approach. With a high Inter-Annotator Agreement (IAA) score of $\alpha = 0.72$, the automated approach shows **substantial agreement with human annotators**, highlighting its potential for consistent and reliable annotations. However, it still lags behind the manual approach in terms of annotation quality, as reflected in the IAA score of manual annotation ($\alpha = 0.87$). The **trade-off between annotation quality and time** is evident, with the Automated Annotation Module completing the task in approximately 6 minutes compared to the 600 minutes required for manual annotation.

While the automated module is efficient, it has limitations such as **missing implicit context** and **annotating turns containing only premises without claims**, often leading to disagreements with human annotators. Attempts to improve GPT-4's performance by providing the annotation manual as input were unsuccessful, as the model performed worse due to multitasking issues. Future research should explore ways to enhance the automated approach, such as incorporating step-by-step instructions from the annotation manual or providing a chain of thought to improve the model's performance.

5.3 Data Augmentation Module

The key advantage of a Data Augmentation approach is the ability to produce large volumes of data quickly and at a lower cost compared to manual data collection, annotation, and labeling. Our Data Augmentation Module aims to improve the generalization capabilities of our Automated Annotation Module by artificially increasing the diversity of our dataset. In this section, we will provide some insights on the data generated by GPT-4 in Subsection 5.3.1 and then evaluate the accuracy and consistency of GPT-4, compared to human annotators, in annotating this data generated in Subsection 5.3.2.

5.3.1 Data Generation Finding and Results

GPT-4 generated a total of 533 turns. For future studies, we intend to conduct a more thorough qualitative analysis of the synthetic data generated by GPT and compare it with the real data produced by the students. However, we have already identified some important points:

- **The GPT's ability to assume roles:** In some discussion rooms where the topic was racism, GPT assumed a minority role (identifying as Latino) and even described experiences involving racism. This impressive level of interaction warrants further study in future research.

- **Dependence on moderator questions:** We imposed a dependence on the questions proposed by the moderator in the prompts of synthetic data, which closely mirrored the real data. Conversations among Portuguese secondary school students exhibit a high degree of dependence on the moderator, responding only to what is asked. By instructing GPT similarly, it managed to replicate this behavior seen in the real data.

- **Degree of formality and absence of errors:** GPT writes almost perfectly, with correct punctuation and no mistakes. In contrast, the real data is filled with emojis and abbreviations.

Future studies should explore GPT's ability to generate data while simulating a high school student to understand if it can emulate these informal behaviors.

- **Lack of justifications:** A notable characteristic of student-generated data is the lack of arguments. There are numerous claims but almost no premises. Even when GPT was asked to write responses with both claims and premises, it mostly produced responses with claims only. This mirrors the student data and suggests that this might be a characteristic of the medium itself or indicative of the type of conversations used in GPT training.

5.3.2 Data Annotation Finding and Results

We evaluated the GPT-4 model on the generated dataset by using a zero-shot prompt approach, instructing GPT to sequentially generate the data and annotate it. This evaluation was done by employing the Methodology 1 described in Subsection 5.1.1 which achieved an **Inter-Annotator Agreement (IAA) score α = 0.57** between the human annotators and the model.

Furthermore, we utilized the annotations from three annotators—two human and the model—to create a filtered dataset. We selected only the majority data, which includes only the data where the human annotators agree, and consolidated it into a single list. This filtering process ensured that in the model's annotations, we retained only those instances that had unanimous human agreement. Consequently, we could evaluate the model's performance by calculating the F1 score, precision, recall, and accuracy, using the majority data from human annotators as the true labels and the model's annotations as the predictions.

	IAA	**F1-score**	**Precision**	**Recall**	**Accuracy**
Real dataset	0.72	0.87	0.84	0.89	0.86
Generated dataset	0.57	0.72	0.62	0.86	0.74

Table 5.3: Comparison of evaluation of GPT's perfomance on the real dataset versus the generated dataset.

As shown in the table, GPT performance declined substantially when annotating the generated dataset, in contrast to the results obtained from annotating the real dataset. We believe this is due to the complexity and volume of information being process at once in one single prompt, **overloading GPT's performance**. Managing generating and subsequently annotating the generated data at once can lead to issues such as prompt fatigue, where the model's attention and performance degrade over the course of processing a long, uninterrupted sequence, thus, lowering the quality of the annotations. Handling both tasks at the same time could overwhelm the model, leading to reduced performance on both fronts. Future research should explore ways to enhance the automated annotation module in annotating generated data. For instance, we could employ a few shots prompt approach with real data examples from our dataset or by providing a Chain of Thought with step-by-step instructions in separate prompts: one for generating the data and another subsequent one to annotate the generated data. The latter option could mitigate overloading GPT's performance by providing separate prompts, hence, improving its accuracy in annotating the data.

5.3.3 Conclusion

In terms of data generation, GPT-4 revealed that it can convincingly assume roles, such as identifying with minority groups, and the data generated mirrored real discussions in its dependence on moderator questions. However, GPT-4's text is formal and error-free, in contrast to the informal, error-prone real student data. Both generated and real data lacked argumentative justifications, indicating a characteristic of the conversation medium.

In terms of data annotation, the Inter-Annotator Agreement (IAA) score was lower for the generated dataset (0.57) compared to the real dataset (0.72). GPT-4 performed better on real data across F1 score, precision, recall, and accuracy. Combining data generation and annotation in one prompt led to performance decline, likely due to overloading GPT-4's processing capabilities.

Future research should focus on improving automated annotation by using few-shot learning and separating data generation and annotation into distinct prompts to enhance GPT-4's performance.

6

PLANNING

Note: The calendar image is available on the project's GitHub[1], as it is too large to place in this report in a compatible way.

The work schedule for this research project adopts Agile Methodology tools, utilizing the Gantt Chart as the primary planning tool. The calendar outlines our planning from the beginning of the project until its completion, highlighting our main key tasks and milestones. Each phase is carefully planned to ensure realistic and achievable goals, contributing to the project's overall success.

The calendar is broken down on a weekly basis. The tasks have been divided into phases, with the start and end dates being relative to the deliveries. Progress is also indicated in each task bar and in the "Percent Complete" column.

- Project Phases:

  1. **Initial Planning and Setup:**
     a) Task Definition;
     b) Tools and Technology Setup;
     c) Preliminary Literature Review.
  2. **Data Annotation, Development, and Testing:**
     a) Manual Annotation;
     b) Development of Annotation Manual and Guidelines;
     c) Prompt Engineering and Development of Automated Annotation Module;
     d) Integration of Data Augmentation Techniques;
     e) Testing and Validation.
  3. **Analysis and Reporting:**
     a) Data Analysis;
     b) Performance Evaluation;
     c) Report Writing and Revision.

[1] https://github.com/DEISI-ULHT-TFC-2023-24/TFC-DEISI89-GPTArgMine

6.1 Challenges and Adjustments

During the course of the project, we made significant changes in the scope of work, which impacted our timeline. These changes were necessary to address newly discovered insights and ensure the effectiveness of our methodology. However, they resulted in delays, particularly in the report writing phase. Despite these challenges, the project was successfully concluded with all objectives met.

The initial scope focused on comparing supervised and unsupervised approaches for argumentation mining in multi-party dialogues. As the project progressed, it became evident that a more fundamental question needed to be addressed: **Can a Machine Learning model perform argumentation mining in multi-party dialogues tasks?** This shift in focus required additional rounds of literature review, and annotation, prompt engineering and testing.

To manage the delays caused by scope changes, we prioritized critical tasks to ensure the most impactful aspects of the project were completed first. Additionally, regular reviews and meetings were conducted weekly to assess development and report progress. And, if necessary, make adjustments to the plan.

7

Conclusion and Future Work

The research undertaken in this thesis highlights the potential and challenges associated with using Large Language Models (LLMs) like GPT-4 for argumentation mining in multi-party dialogues. Our work focused on leveraging these models to identify argumentative language in conversations among Portuguese high school students. The study introduced a novel approach to annotation by comparing manual and automated methods, supplemented by data augmentation techniques.

7.1 Key Findings

- **Manual Annotation:** Achieved a high Inter-Annotator Agreement (IAA) score, validating the reliability of the annotation guidelines. Despite its accuracy, manual annotation proved to be time-consuming and labor-intensive.
- **Automated Annotation:** The GPT-4 model demonstrated substantial agreement with human annotators, showcasing its potential as an efficient and consistent annotator. However, the IAA score was lower than manual annotations, indicating room for improvement.
- **Data Augmentation:** The synthetic data generated by GPT-4, while valuable in expanding the dataset, exhibited limitations such as a lack of informal language and justifications typical of real student-generated data.

This thesis underscores the significant progress made in automating the annotation process but also highlights the challenges, particularly in maintaining high annotation quality and accurately reflecting real-world communication nuances.

7.2 Future Work

To build on the findings of this thesis and address the identified challenges, several key points are proposed for future research and development. By pursuing these directions, future research can improve the accuracy, efficiency, and applicability of LLMs in argumentation mining, ultimately contributing to more sophisticated and reliable NLP applications.

- **Enhancement of Automated Annotation:** Implement a few-shot prompt approach with real data examples to improve the accuracy of GPT-4 annotations. Develop a Chain of Thought approach with step-by-step instructions to guide the model in generating and annotating data separately.
- **Improvement in Data Augmentation:** Conduct qualitative analyses of synthetic data to better understand the gaps compared to real data. Explore techniques to induce informal language and justifications in synthetic data to better mimic real student conversations.
- **Evaluation of Different LLMs:** Test and compare the performance of various LLMs, beyond GPT-4, for both annotation and data generation tasks.Investigate the impact of model fine-tuning on annotation quality.
- **Hybrid Annotation Approaches:** Develop hybrid models that combine automated and manual annotation techniques to leverage the strengths of both methods. Explore active learning strategies, where the model suggests annotations that are then reviewed by human annotators.

BIBLIOGRAPHY

van Eemeren, F. H., Garssen, B., Krabbe, E. C., Henkemans, A. F. S., Verheij, B., & Wagemans, J. H. (2014). *Handbook of argumentation theory*. https://doi.org/10.1007/978-90-481-9473-5 (cit. on pp. 1, 7).

Modgil, S., Toni, F., Bex, F., Bratko, I., Chesnevar, C. I., Dvořák, W., Falappa, M. A., Fan, X., Gaggl, S. A., García, A. J., et al. (2013). The added value of argumentation. *Agreement technologies*, 357–403 (cit. on p. 1).

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, *45*(4), 765–818 (cit. on p. 1).

Mochales, R., & Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, *19*, 1–22 (cit. on p. 2).

Teruel, M., Cardellino, C., Cardellino, F., Alemany, L. A., & Villata, S. (2018). Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (cit. on p. 2).

Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R., & King, J. (2012). A corpus for research on deliberation and debate. *LREC*, *12*, 812–817 (cit. on p. 2).

Lippi, M., & Torroni, P. (2016). Argument mining from speech: Detecting claims in political debates. *Proceedings of the AAAI conference on artificial intelligence*, *30*(1) (cit. on p. 2).

Abbas, S., & Sawamura, H. (2011). Ales: An innovative agent-based learning environment to teach argumentation. *International journal of knowledge-based and intelligent engineering systems*, *15*(1), 25–41 (cit. on p. 2).

Reshamwala, A., Mishra, D., & Pawar, P. (2013). Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ)*, *3*(1), 113–116 (cit. on p. 2).

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258 (cit. on p. 2).

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075* (cit. on pp. 2, 3).

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* (cit. on p. 2).

Ding, B., Liu, L., Bing, L., Kruengkrai, C., Nguyen, T. H., Joty, S., Si, L., & Miao, C. (2020). Daga: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549* (cit. on p. 2).

Liu, L., Ding, B., Bing, L., Joty, S., Si, L., & Miao, C. (2021). Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. *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)*, 5834–5846 (cit. on p. 3).

Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487* (cit. on p. 3).

Ding, B., Qin, C., Liu, L., Chia, Y. K., Joty, S., Li, B., & Bing, L. (2022). Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450* (cit. on pp. 3, 16, 17).

Cabrio, E., & Villata, S. (2018). Five years of argument mining: A data-driven analysis. *IJCAI*, *18*, 5427–5433 (cit. on p. 3).

Dusmanu, M., Cabrio, E., & Villata, S. (2017). Argument mining on twitter: Arguments, facts and sources. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2317–2322 (cit. on p. 3).

Dignum, F. P., & Vreeswijk, G. A. (2003). Towards a testbed for multi-party dialogues. *Workshop on Agent Communication Languages*, 212–230 (cit. on p. 4).

Traum, D. (2003). Issues in multiparty dialogues. *Workshop on Agent Communication Languages*, 201–211 (cit. on p. 4).

Rocha, G., Cardoso, H. L., & Teixeira, J. (2016). Argmine: A framework for argumentation mining. *Computational Processing of the Portuguese Language-12th International Conference, PROPOR*, *13* (cit. on p. 7).

van Eemeren, F. H., Houtlosser, P., & Henkemans, A. F. S. (2007). *Argumentative indicators in discourse: A pragma-dialectical study*. Springer. (Cit. on p. 8).

Perelman, C., & Olbrechts-Tyteca, L. (1969). The new rhetoric: A treatise on argumentation. 1958. *Trans. John Wilkinson and Purcell Weaver. Notre Dame: U of Notre Dame P* (cit. on p. 8).

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, 40–49 (cit. on p. 9).

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189 (cit. on p. 9).

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, *34*(4), 555–596 (cit. on p. 9).

Wacholder, N., Muresan, S., Ghosh, D., & Aakhus, M. (2014). Annotating multiparty discourse: Challenges for agreement metrics. *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, 120–128 (cit. on p. 9).

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability (cit. on p. 9).

Stab, C., & Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, 1501–1510 (cit. on p. 9).

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107 (cit. on p. 10).

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160 (cit. on p. 11).

Trautmann, D., Daxenberger, J., Stab, C., Schütze, H., & Gurevych, I. (2020). Fine-grained argument unit recognition and classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9048–9056 (cit. on p. 11).

Persing, I., & Ng, V. (2020). Unsupervised argumentation mining in student essays. *Proceedings of the 12th Language Resources and Evaluation Conference*, 6795–6803 (cit. on p. 11).

Eger, S., Daxenberger, J., & Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104* (cit. on p. 11).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9 (cit. on p. 13).

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., et al. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* (cit. on p. 13).

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (cit. on p. 13).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837 (cit. on pp. 13, 14).

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30* (cit. on p. 14).

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99 (cit. on p. 15).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press. (Cit. on p. 15).