# DEBISS-Arg: An In Depth Data Annotation Protocol and Corpus for Argument Mining in Semi Structured Debates

**David Eduardo Pereira**[1], **Daniela Thuaslar Simão**[2], **Claudio E. C. Campelo**[1]

[1]Systems and Computing Department
Federal University of Campina Grande (UFCG)
Campina Grande – PB – Brazil

[2]Languages Department
Federal University of Campina Grande (UFCG)
Campina Grande – PB – Brazil

davidpereira@copin.ufcg.edu.br, daniela.thuaslar@estudante.ufcg.edu.br,

campelo@dsc.ufcg.edu.br

***Abstract.*** *Argumentation plays a crucial role across various domains of human activity. However, its diverse applications present challenges in knowledge representation, leading to the development of numerous data models without a universally accepted standard. This lack of standardization complicates the creation of data and representation frameworks for argument annotation, which are essential for building high-quality datasets. This research addresses these limitations by proposing a comprehensive data annotation protocol specifically designed for argument mining in semi-structured debates. The protocol is applied to the newly introduced DEBISS-Arg corpus, which includes multiple annotation labels covering a range of argument mining tasks in Brazilian Portuguese.*

## 1. Introduction

The process of argumentation is present in our lives and proves to be indispensable for a variety of activities such as scientific research, essay writing, debates, legal trials, and even everyday tasks like defending a point of view or discussing ideas. Moreover, argumentation stimulates investigation by encouraging the search for facts and evidence to support conclusions, which aids in the development of scientific, creative, and critical thinking.

However, representing arguments presents significant challenges in knowledge representation due to the absence of a universal analytical system, as arguments vary greatly depending on context and can also be influenced by social, cultural, linguistic, and rhetorical factors, as well as the personal style and preferences of the writer. Typically, an argument is composed of claims and premises [Kotelnikov et al. 2022]; it can also include other components such as counterclaims and evidence to strengthen its persuasiveness. Researchers conducted a qualitative analysis across six different datasets and found that these datasets conceptualize claims, which represent a stance on a topic, quite differently, as each has its own definition [Daxenberger et al. 2017]. Consequently, developing a standardized system that can account for all possible variations and nuances of argumentation is a challenging task [Gao 2024].

Given these challenges and the advancement of AI, a new area of research has garnered attention: Argument Mining (AM). This field involves the automatic identification of text structures expressing an argument and its respective analysis, within Natural Language Processing (NLP). Since approximately 2014 [Cabrio and Villata 2018], AM research has focused on automatically identifying and analyzing argumentative structures in diverse texts, including essays, [Sazid and Mercer 2022, Wambsganss et al. 2020a, Abkenar et al. 2021, Guo et al. 2024], scientific papers [Michael, Fromm et al. 2020, Al Khatib et al. 2021, Stylianou and Vlahavas 2021, Accuosto et al. 2021, Fergadis et al. 2021, Binder et al. 2022, Wang et al. 2024], legal texts [Westermann et al. 2022, Zhang et al. 2022], news articles [Lavee et al. 2019], among others.

Beyond written texts, AM extends to audio transcriptions, including political debates [Duthie et al. 2016, Mestre et al. 2021, Mancini et al. 2022], podcasts [Pojoni et al. 2023], and professional debates [Mirkin et al. 2018]. Additionally, researchers explore techniques for the automatic generation of arguments. For instance, IBM's Debater project aims to develop artificial intelligence capable of engaging in debates with humans[Bar-Haim et al. 2021]. In the context of AM, studies address both written debates on online platforms [Sousa et al. 2021, Habernal and Gurevych 2016, Boltužić and Šnajder 2016, Chakrabarty et al. 2019], such as Twitter and Reddit, and oral debates, with political debates receiving significant attention. Numerous established datasets provide detailed annotations on argumentation at various levels for political debates [Haddadan et al. 2019, Lippi and Torroni 2016]. These debates are usually highly structured, meaning that they have a limited number of rounds and speeches, as well as time constraints.

When considering AM research in the field of debates, it is noteworthy the scarcity of studies that diverge from the typical format of political and highly structured debates. Specifically, research on oral, individual, and semi-structured debates — such as those found in academic settings, university conferences, or daily discussions- remains under explored in the literature. There is a particular scarcity of studies focusing on the educational environment and the development of oratory skills in student debates. This presents a critical gap in the AM literature, as these semi-structured interactions offer unique argumentative dynamics that differ significantly from highly formalized debates.

Addressing this critical gap, this research introduces the DEBISS-Arg corpus and presents a comprehensive AM protocol specifically designed for it. This protocol is built upon existing annotation schemes and theoretical definitions, but crucially adapts and extends them to capture the unique challenges of spoken debates with individual and semi-structured characteristics, an area significantly underexplored in current AM research. To support this effort, a new corpus of debate transcriptions, characterized by individual, spoken, and semi-structured features (DEBISS corpus [Souza et al. 2025]) was developed and is employed in this study as the foundation for applying and evaluating the proposed annotation protocol. The data is available on GitHub page[1].

[1]https://github.com/AINDA-Project-UFCG/argument-mining-data

## 2. Related Work

Argumentation has been extensively studied across various fields since ancient Greek philosophy [Bentahar et al. 2010]. However, a significant challenge persists: there is no universal perspective or standard for argumentation models and methods , which complicates the selection of suitable frameworks for new intelligent software systems. [Walton et al. 2008, van Eemeren et al. 2014]. Additionally, AM approaches struggle due to a lack of large, well-annotated datasets; nevertheless, recent efforts are being made to create corpora across various domains [Lawrence and Reed 2020]. Consequently, this section delves into how AM researchers define arguments, their components, and the relationships required for data modeling and annotation. By reviewing key existing models, we aim to establish a clear foundation for the annotation protocol proposed in this research, emphasizing how it leverages and adapts established concepts to address specific contextual needs.

Argumentation schemes are attracting attention from those interested in argumentation and AI, highlighting the interdisciplinary nature of the research field [Reed and Norman 2003]. While efforts to formalize and compile these schemes exist [Walton et al. 2008], a persistent challenge is the lack of universal consensus. To address this, researchers have proposed categorizing argument models into three main structures: rhetorical, dialogical, and monological (Figura 1). This classification provides a framework for understanding and selecting appropriate models for specific contexts [Bentahar et al. 2010] which is particularly relevant for the design of comprehensive annotation protocols.

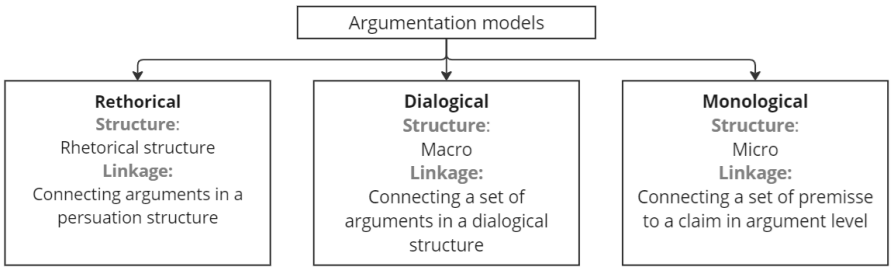



**Figure 1. Taxonomy proposed by [Bentahar et al. 2010]**

According to this research, **monological** models are defined as those that focus on the internal structure (micro) of an argument and the relationships within its elements, rather than the relationships between different arguments. Meanwhile, **dialogical** models analyze exchanges between multiple parties or speakers, examining the interactions (macro) [Bentahar et al. 2010, Habernal and Gurevych 2017]. Completing this proposed taxonomy, rhetorical models are the final classification. The main idea behind rhetorical models is to consider the audience's perspective on the argumentative discourse, focusing on its rhetorical persuasiveness. These models aim to establish an evaluative view of the propositions made by arguments, rather than focusing on the micro or macro structures of arguments.

However, these models are not completely distinct; they can share similarities. Consequently, some models exhibit characteristics of both classification. For example, there are models that combine monological and dialogical structures of arguments

[Bentahar et al. 2010]. Despite their distinct categorizations, it is important to remember that these models can overlap and be combined to create more precise and effective approaches. By leveraging the strengths of different models and integrating them strategically, researchers can develop comprehensive methods that better capture the complexities of argument data representation and analysis.

Researchers have compiled a compendium of about 60 different types of argument schemes [Walton et al. 2008], which includes defined rules, examples, and references for each model. This research highlights the application of these models to AI, showcasing tools for representing arguments according to the proposed schemas. While not exhaustive, this compendium illustrates the inherent complexity of the field. A foundational model widely applied across various discourse types, including daily written and spoken arguments, is the Toulmin model. Proposed by Stephen Toulmin in 1958, this model describes the microstructure of an argument through components like claim, data (or ground), warrant, qualifier, rebuttal, and backing.

Argument models have profoundly influenced AM research, particularly in AI solutions for debates. Although progress has been made, there are still some gaps that need to be addressed. For example, the research conducted in [Duthie et al. 2016] applies AM using AI solutions to analyze political debates in US elections. This study focuses on identifying **ethos** (appeal to the credibility or authority of the speaker), a common form of argumentative structure found in political debates. However, other research that performs micro-level analysis of arguments goes beyond ethos. It includes the identification of Argumentative Discourse Units (ADUs) and their components, such as claims, evidence, and premises [Lippi and Torroni 2016, Mancini et al. 2022, Haddadan et al. 2019].

To understand the macro structures of arguments in debates, methodologies have leveraged argumentation models for AI solutions. Studies such as [Hautli-Janisz et al. 2022, Visser et al. 2019] use argumentation models for macro analysis based on Anchor Theory, proposed by [Budzynska and Reed 2011], which is grounded in the Speech Act Theory proposed by Austin in 1962 and Searle in 1969.

Beyond formal debates, argumentation mining extends to social media. Research on Twitter, for instance, treats threads as debate structures to identify components like premises and stance [Bhatti et al. 2021, Feger and Dietze 2024, Chakrabarty et al. 2020]. In this context, each tweet is considered an ADU where premises are extracted and hashtags indicate stance.

Overall political debates — typically spoken and based on data extracted from transcriptions — are the most common context used by researchers. Another context identified is online and written debates, with data sourced from platforms like Twitter or Reddit. However, as previously mentioned in this article, research on debates remains limited. Consequently, the current research aims to investigate existing approaches and develop new annotation protocols to advance the field of AM in debates, which remains underexplored in the state of the art.

## 3. Data and Annotation Protocol

This section presents the proposed annotation protocol (Section 3.2), along with a brief description of DEBISS corpus (Section 3.1). The protocol was designed to support in-depth AM analysis of debates that exhibit characteristics similar to those found in the

DEBISS corpus. Following the definition of the corpus and the annotation protocol, this section also details the annotation process for the DEBISS-Arg corpus and outlines the methodology adopted for annotation and review in Section 3.3.

### 3.1. The DEBISS Corpus

The corpus consists of 9 hours and 35 minutes of audio literal transcriptions from in-person debates held in Brazilian Portuguese. These debates were conducted in 2024 with 67 first-year computer science undergraduate students, organized into 16 groups. A moderator facilitated the sessions, explaining rules, managing discussions, and promoting interaction. All participants signed informed consent forms authorizing the use of the data for research purposes. To ensure anonymity, each debater was randomly assigned a number and referred to as "Debater 1," "Debater 2," and so forth, according to the number of participants in each group. With this procedure, no personal data will be available in the published dataset since no personal identifiers were included in the transcripts.

The debates, facilitated by a moderator who explains the rules, maintains order, promotes interaction, and ensures equitable participation, are conducted in a semi-structured format that blends predefined mandatory questions with opportunities for unrestricted expression. This flexible approach allows participants to freely present viewpoints, respond, and counter-argue without interruptions, following a multi-stage structure organized into three distinct parts: first, debaters share their initial opinions; second, they address targeted questions from the moderator, engaging in a broader exchange enriched by peers' comments and questions; and third, a final question and mandatory reflections prompt participants to reassess their initial positions, offering valuable insights into the persuasive dynamics of the discussion.

This corpus is designed not only to meet the needs of AM tasks but also to include rich annotations that can be used in various contexts for NLP tasks. The corpus includes annotations for disfluency detection (DEBISS-Disfluecy [Lima and Campelo 2024]), debater quality analysis (DEBISS-Eval[2]), and AM tasks (DEBISS-Arg). The corpus can also be highly useful for speech-to-text tasks, voice print, speaker diarisation, providing a significant advancement in the availability of open data for NLP tasks in **Brazilian Portuguese**. Moreover, the transcriptions were kept literal to preserve a faithful representation of spoken debates, including stuttering, hesitations, repetitions, and speech disfluencies. For more details on the DEBISS corpus data collection, methodology and consent procedures, readers may refer to the study presented in [Souza et al. 2025]. The dataset is available on the GitHub page[3].

### 3.2. Proposed protocol

The main goal of the proposed protocol is to conduct an in-depth analysis of debates close to the DEBISS corpus format, embracing the diverse characteristics present in these kind of debates. It is important to note that this protocol does not propose a new argumentation model or schema. Instead, it consolidates several definitions that were validated in the process of AM by research in various scenarios and applications and are adequate for the proposed context. It leverages the strengths of these models to adapt them for the context

[2]https://github.com/AINDA-Project-UFCG/debater-quality-data
[3]https://github.com/AINDA-Project-UFCG/transcription-data

of debates in the given format. To build the proposed protocol, it was necessary to conduct an in-depth investigation of the existing models, aiming to identify the theoretical aspects on which those models are based, as well as the methodology for data annotation that was used.

The first and most basic process adopted by many data protocols is the identification of Argumentation Units (AU), also known ADU [Fergadis et al. 2021, Sazid and Mercer 2022, Westermann et al. 2022]. Essentially, these protocols determine whether a given sentence is considered an ADU or non-ADU. The identification of ADUs is the primary step in many data argumentation protocols. However, the definition of an ADU is not universal and can vary depending on the context. For example, in Twitter research, an ADU is often defined as the entire text of a tweet [Bhatti et al. 2021]. According to Walton's definition, an ADU should include a set of premises and a conclusion, also the author defines numerous forms of acceptable arguments with specific schemas [Walton et al. 2008].

Conversely, some consider an argument to be an entire utterance of a dialogue that contains some argumentative structure in the text [Hautli-Janisz et al. 2022], particularly when considering dialogical models, as is the case for this protocol. Upon analyzing the actual data, it was observed that each utterance may contain more than one argument. For this reason, each ADU is contained within a single utterance; however, a single utterance might contain one or more arguments or none at all if it lacks an argumentative structure. In some instances, a single utterance might serve as a single ADU if the entire utterance forms a cohesive argumentative structure.

The identification of ADUs takes into consideration the macro structure of arguments in a dialogical format. However, the proposed protocol aims to conduct an in-depth analysis, which necessitates the inclusion of microstructure definitions that compose an argument. One of the primary components is the claim [Lippi and Torroni 2016, Haddadan et al. 2019, Abkenar et al. 2021], a concept widely adopted by researchers and featured in the Toulmin classical model. A claim is essentially an affirmation or proposition, often referred to as an opinion in debates. It is typically a straightforward and explicit proposition. This component is mandatory in an argument, and every ADU in this protocol must include a claim, consistent with similar protocols proposed by other researchers [Fergadis et al. 2021, Sazid and Mercer 2022, Kotelnikov et al. 2022].

Moreover, another important component of an argument is evidence, which is also present in Toulmin's model and supported by research across different contexts [Lippi and Torroni 2016, Mancini et al. 2022, Stylianou and Vlahavas 2021]. Evidence refers to a span of text within an argument that provides information used to support or challenge a given claim or premise. Evidence may have subtypes, which are particularly useful in identifying this component within an argument. After reviewing the available classifications proposed in the literature and examining examples of evidence in the data, the following classes were identified and incorporated into this protocol: Historical Context, Examples, Data, Expert Opinions (citation), Debater Mention, and Common Sense (Face Value), these defined types are based on researcher that adopts a similar approach with some different group of classes [Aharoni et al. 2014, Lippi and Torroni 2016].

Also, considering the microstructure of an argument, another relevant compo-

nent is the premise. A premise is a certainty, belief, statement, or proposition about the subject that is used to develop an argument supporting or refuting a claim. An important aspect of a premise is that it may contain evidence. In Walton's framework, premises are defined as major and minor. For this protocol, we will consider all premises without differentiating between them. Premises are also commonly used in various contexts within AM [Bhatti et al. 2021, Haddadan et al. 2019, Sazid and Mercer 2022, Abkenar et al. 2021, Habernal and Gurevych 2015].

Another important aspect of argumentation, in addition to the ADU and its components, is the relationships among these components. These relationships can be classified into micro and macro structures of the debate [Bentahar et al. 2010]. In the microstructure, we define the relationships among the components of a single argument. In contrast, the macrostructure involves the relationships among different ADUs, or typically different utterances in dialogical models. This protocol is particularly well-suited for dialogical models involving multiple parties or individuals discussing a topic. Consequently, understanding the relationships among arguments or different utterances is crucial for a thorough analysis of the debates' characteristics. The macro relationships framework adopted in this research is inspired by the Anchor Theory. We incorporated key aspects of this model that are well-suited to the goals of this protocol.

Taking into account the Anchor Theory definition and the data format, the most suitable classifications for macro relationships in this protocol are: questioning, agreement, partial agreement, disagreement, and reference. These categories collectively map various aspects and interactions among debaters, enriching the debate analysis. For micro structures, we focus on two types of relationships: support and attack [Mestre et al. 2021, Wambsganss et al. 2020b, Zhang et al. 2023]. These are commonly addressed in research studies that explore relational aspects at the micro level and are used to map the relationship between premises and evidence to claims.

In Figure 2, it is presented an illustration of a hypothetical debate involving three debaters and a moderator who manages the discussion. The figure display the identification of ADUs and non-ADUs (note that while this simplified version shows each utterance containing a single ADU, in practice, there may be multiple ADUs within an utterance). The illustration highlights microstructures such as premises, claims, and evidence, and also depicts macrostructural relationships like disagreement, agreement, and questioning. Additionally, the micro-relations among components of each ADU, including support and attack, are illustrated. Table 1 lists all the available annotations defined in the proposed protocol.

### 3.3. DEBISS-Arg Corpus

**Annotation process:** the annotation protocol was shared with professional annotators who have a degree in linguistics and professional experience in labeling text for a variety of NLP tasks, ensuring they were well-suited for the manual annotation of the DEBISS-arg corpus. Two professional annotators participated in the process, each responsible for labeling half of the dataset using the tailored protocol. For the text labels annotation process, a proprietary web-based annotation tool (Figure 3) was used, providing annotators with access to the transcription text. Within this tool, they could highlight spans of text and assign appropriate labels based on the categories defined in the protocol. Figure 3 presents a screenshot of the tool, illustrating how the annotation process was carried out.

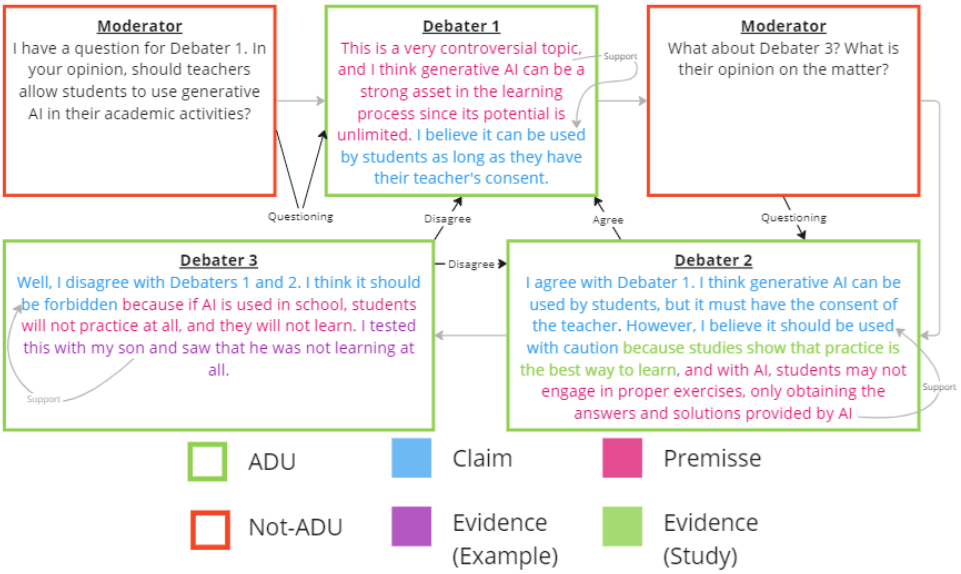



**Figure 2. Protocol usage illustration of a simplified debate**

**Table 1. DEBISS-Arg Labels**

| Label | Type |
|---|---|
| ADU, Premise, Claim, and Evidence (example, expert opinion, history, context, debater citation, data) | Plain Text |
| Support and Attack | Relation Label Micro Level |
| Questioning, Agree, Disagree, Argue and Partial Disagree | Relation Label Macro Level |

Furthermore, beyond the text span annotation process carried out in the web tool — which includes all the labels in Table 1, from the most basic ones such as ADUs to the argumentative components — an additional layer of annotation was performed. This accounted for both micro and macro relations. However, since the proprietary annotation tool used for text span labeling does not support linking text spans or specifying the type of relationship between them, an alternative approach was adopted using Google Spreadsheets. These spreadsheets were later exported to CSV files, which are available on the corpus GitHub page.

**Data reviewing process:** at the end of the annotation phase, a pairwise revision process was performed. During this stage, each annotator reviewed the labels assigned by the other. When disagreements were identified, the reviewer and the original annotator collaborated to reach a final consensus on the annotations. Hence, to better understand the nature of these disagreements, a categorization scheme was created to describe the changes made from the original annotation to the version obtained after the revision process. These changes are classified as follows:

- Add Label: When the annotator adds a new label that was not present in the text before.

**Figure 3. Web Based Annotation Tool**

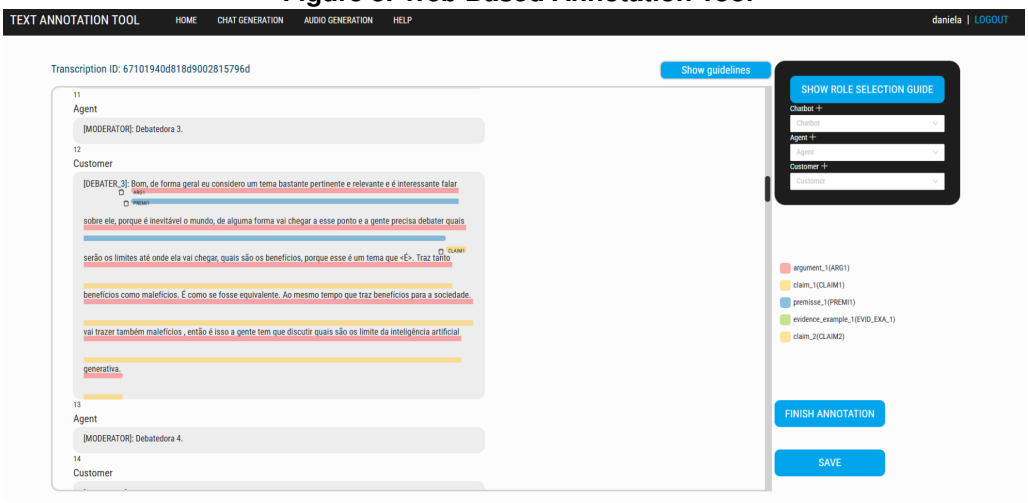
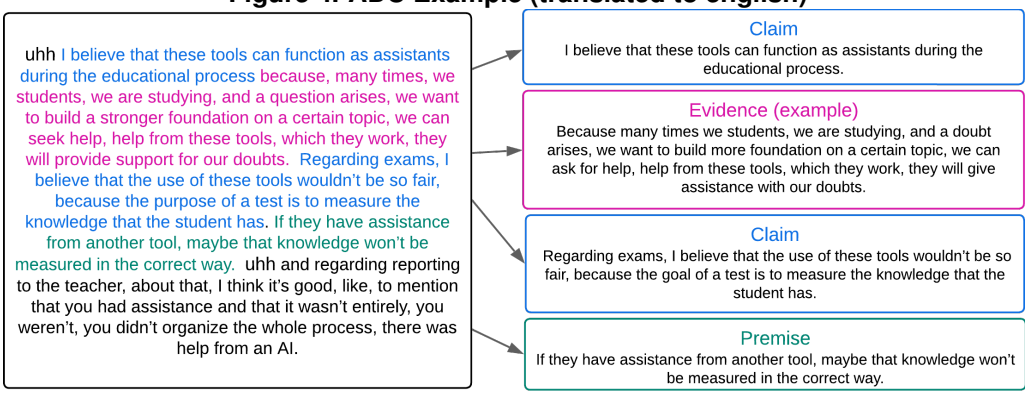



**Figure 4. ADU Example (translated to english)**




- Change Label: When a text span was assigned a label and is later changed to a new one.
- Change Label Text: When the label remains the same, but the text span is modified in some way, such as by adding or excluding tokens from the sentence.

Additionally, Levenshtein metrics — used to measure textual similarity — were calculated between the original annotations and the final annotations established through discussion and consensus between the annotators. These metrics help quantify the extent of modifications made during the revision process, offering insights into the consistency of the annotation workflow.

Out of 549 total utterances, 91 were modified, resulting in a total of 70 **add label**, 28 **change label**, and 12 **change label text**. The median similarity value was 60%, with the highest being 91% and the lowest 16%. For detailed metric information across each debate section, a comprehensive table is available on the project's GitHub page[4]. These mapped changes help illustrate the nature of the adjustments made and offer valuable insights into the annotation process conducted by professional annotators.

[4]https://github.com/AINDA-Project-UFCG/argument-mining-data/blob/main/revision-metrics.png

**Table 2. Stats for DEBISS-Arg Corpus**

| Type | Label Name | Amount | Tokens | Mean Tokens | MedianTokens |
|---|---|---|---|---|---|
| Text Span | Argument | 210 | 156152 | 134.76 | 102 |
| | Premise | 287 | 34400 | 35.87 | 30 |
| | Claim | 384 | 59760 | 28.88 | 26 |
| | Evidence | 323 | 68434 | 39.92 | 30 |
| Micro | Support | 336 | - | | |
| | Attack | 19 | - | | |
| Macro | Question | 130 | - | | |
| | Agree | 65 | - | | |
| | Disagree | 26 | - | | |
| | Partially Agree | 15 | - | | |
| | Discussion | 5 | - | | |

**DEBISS-Arg corpus stats:** the final result of the DEBISS-Arg corpus are summarized in Table 2, where we can find the amount of labeled text spans for each of the ten categories, as well as token and character counts. Additionally, the table presents the number of relation annotations for both micro and macro structures. The DEBISS-Arg is a comprehensive AM corpus in Brazilian Portuguese and represents one of the important state of the art contribution, as it can support various AM tasks. An ADU example, illustrating the data, can be found in Figure 4. This example was retrieved from the DEBISS-Arg corpus and translated into English.

## 4. Conclusion

The proposed protocol has been designed to address the unique characteristics of arguments found in DEBISS corpus. The protocol's primary goal is to facilitate a comprehensive analysis of debates by incorporating both micro and macro structures of argumentation. By distinguishing between ADUs and their components — such as claims, evidence, and premises — the protocol builds upon established models and definitions while adapting them to the specific context of the debates. The protocol also emphasizes the importance of understanding relationships among arguments at both micro and macro levels, utilizing classifications like support, attack, questioning, agreement, and disagreement to map the interactions among debaters. The integration of these elements, inspired by the Argumentation Anchor Theory, enhances the depth and quality of debate analysis.

Furthermore, this research introduces the DEBISS-Arg corpus, which contains a practical application of the proposed protocol, through the annotation and evaluation process carried out by professional annotators. This is an important contribution, given the scarcity of available open source data for various NLP tasks, particularly those focused on AM. Overall, the protocol and corpus provides a comprehensive framework for exploring and evaluating argumentation within structured debate environments, offering valuable insights into the dynamics of argumentative discourse. Although the protocol was applied specifically to the presented corpus, it shows strong potential for generalization and could be adapted for use in other debate scenarios and corpora in future work.

## References


Abkenar, M. Y., Stede, M., and Oepen, S. (2021). Neural argumentation mining on essays and microtexts with contextualized word embeddings (short paper). In *Swiss Text Analytics Conference*.

Accuosto, P., Neves, M. L., and Saggion, H. (2021). Argumentation mining in scientific literature: From computational linguistics to biomedicine. In *BIR@ECIR*.

Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., and Slonim, N. (2014). A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In Green, N., Ashley, K., Litman, D., Reed, C., and Walker, V., editors, *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Al Khatib, K., Ghosal, T., Hou, Y., de Waard, A., and Freitag, D. (2021). Argument mining for scholarly document processing: Taking stock and looking ahead. In Beltagy, I., Cohan, A., Feigenblat, G., Freitag, D., Ghosal, T., Hall, K., Herrmannova, D., Knoth, P., Lo, K., Mayr, P., Patton, R. M., Shmueli-Scheuer, M., de Waard, A., Wang, K., and Wang, L. L., editors, *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics.

Bar-Haim, R., Ein-Dor, L., Orbach, M., Venezian, E., and Slonim, N. (2021). Advances in debating technologies: Building AI that can debate humans. In Chiang, D. and Zhang, M., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Bentahar, J., Moulin, B., and Bélanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Bhatti, M. M. A., Ahmad, A. S., and Park, J. (2021). Argument mining on Twitter: A case study on the planned parenthood debate. In Al-Khatib, K., Hou, Y., and Stede, M., editors, *Proceedings of the 8th Workshop on Argument Mining*, pages 1–11, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Binder, A., Verma, B., and Hennig, L. (2022). Full-text argumentation mining on scientific publications.

Boltužić, F. and Šnajder, J. (2016). Fill the gap! analyzing implicit premises between claims from online debates. In Reed, C., editor, *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany. Association for Computational Linguistics.

Budzynska, K. and Reed, C. (2011). Speech acts of argumentation: inference anchors and peripheral cues in dialogue. In *Proceedings of the 10th AAAI Conference on Computational Models of Natural Argument*, AAAIWS'11-10, page 3–10. AAAI Press.

Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., and Hwang, A. (2019). AMPERSAND: Argument mining for PERSuAsive oNline discussions. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Chakrabarty, T., Hidey, C., Muresan, S., Mckeown, K., and Hwang, A. (2020). Ampersand: Argument mining for persuasive online discussions.

Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (2017). What is the essence of a claim? cross-domain claim identification. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Duthie, R., Budzynska, K., and Reed, C. (2016). *Mining Ethos in Political Debate*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 299–310. IOS Press, Netherlands. This research was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the Polish National Science Centre under grant 2015/18/M/HS1/00620.

Feger, M. and Dietze, S. (2024). Taco – twitter arguments from conversations.

Fergadis, A., Pappas, D., Karamolegkou, A., and Papageorgiou, H. (2021). Argumentation mining in scientific literature for sustainable development. In Al-Khatib, K., Hou, Y., and Stede, M., editors, *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gao, Y. (2024). *Mining Arguments in Scientific Documents*. Doctoral thesis, ETH Zurich, Zurich.

Guo, K., Li, Y., Li, Y., and Chu, S. (2024). Understanding efl students' chatbot-assisted argumentative writing: An activity theory perspective. 29(1):1–20. Publisher Copyright: © 2023, The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature.

Habernal, I. and Gurevych, I. (2015). Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2016). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2017). Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Haddadan, S., Cabrio, E., and Villata, S. (2019). Yes, we can! mining arguments in 50 years of US presidential campaign debates. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Hautli-Janisz, A., Kikteva, Z., Siskou, W., Gorska, K., Becker, R., and Reed, C. (2022). Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3291–3300. European Language Resources Association (ELRA). © European Language Resources Association (ELRA).

Kotelnikov, E., Loukachevitch, N., Nikishina, I., and Panchenko, A. (2022). Ruarg-2022: Argument mining evaluation. In *Computational Linguistics and Intellectual Technologies*. RSUH.

Lavee, T., Orbach, M., Kotlerman, L., Kantor, Y., Gretz, S., Dankin, L., Jacovi, M., Bilu, Y., Aharonov, R., and Slonim, N. (2019). Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. In Stein, B. and Wachsmuth, H., editors, *Proceedings of the 6th Workshop on Argument Mining*, pages 58–66, Florence, Italy. Association for Computational Linguistics.

Lawrence, J. and Reed, C. (2020). Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Lima, P. L. and Campelo, C. E. (2024). Disfluency detection and removal in speech transcriptions via large language models. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 227–235, Porto Alegre, RS, Brasil. SBC.

Lippi, M. and Torroni, P. (2016). Argument mining from speech: Detecting claims in political debates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Mancini, E., Ruggeri, F., Galassi, A., and Torroni, P. (2022). Multimodal argument mining: A case study in political debates. In Lapesa, G., Schneider, J., Jo, Y., and Saha, S., editors, *Proceedings of the 9th Workshop on Argument Mining*, pages 158–170, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Mestre, R., Milicin, R., Middleton, S. E., Ryan, M., Zhu, J., and Norman, T. J. (2021). M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In Al-Khatib, K., Hou, Y., and Stede, M., editors, *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael, Fromm, Max, Berrendorf, Evgeniy, Faerman, and Thomas, Seidl (2020). Argument mining driven analysis of peer-reviews dataset.

Mirkin, S., Jacovi, M., Lavee, T., Kuo, H.-K., Thomas, S., Sager, L., Kotlerman, L., Venezian, E., and Slonim, N. (2018). A recorded debating dataset. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources*

*and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pojoni, M.-L., Dumani, L., and Schenkel, R. (2023). Argument-mining from podcasts using chatgpt. In *ICCBR Workshops*, pages 129–144.

Reed, C. and Norman, T., editors (2003). *Argumentation Machines: New Frontiers in Argument and Computation*. Argumentation Library. Kluwer Academic Publishers, Netherlands.

Sazid, M. T. and Mercer, R. E. (2022). A unified representation and a decoupled deep learning architecture for argumentation mining of students' persuasive essays. In Lapesa, G., Schneider, J., Jo, Y., and Saha, S., editors, *Proceedings of the 9th Workshop on Argument Mining*, pages 74–83, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Sousa, J. P., Nascimento, R., Araujo, R., and Coelho, O. (2021). Não se perca no debate! mineração de argumentação em redes sociais. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 139–150, Porto Alegre, RS, Brasil. SBC.

Souza, K., Pereira, D., and Claúdio, C. (2025). Debiss: a corpus of individual, semi-structured and spoken debates. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC.

Stylianou, N. and Vlahavas, I. (2021). Transformed: End-to-nd transformers for evidence-based medicine and argument mining in medical literature. *Journal of Biomedical Informatics*, 117:103767.

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., and Wagemans, J. H. M. (2014). Handbook of argumentation theory.

Visser, J., Lawrence, J., Wagemans, J., and Reed, C. (2019). An annotated corpus of argument schemes in us election debates. In *Proceedings of the 9th Conference of the International Society for the Study of Argumentation (ISSA), 3-6 July 2018*, pages 1101–1111.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., and Leimeister, J. M. (2020a). Al: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Wambsganss, T., Niklaus, C., Söllner, M., Handschuh, S., and Leimeister, J. M. (2020b). A corpus for argumentative writing support in german.

Wang, S., Zhang, Y., and Du, J. (2024). Utilizing llms to evaluate the argument quality of triples in semmeddb for enhanced understanding of disease mechanisms. *medRxiv*.

Westermann, H., Savelka, J., Walker, V. R., Ashley, K. D., and Benyekhlef, K. (2022). Toward an intelligent tutoring system for argument mining in legal texts.

Zhang, G., Nulty, P., and Lillis, D. (2022). Enhancing legal argument mining with domain pre-training and neural networks.

Zhang, G., Nulty, P., and Lillis, D. (2023). Argument mining with graph representation learning. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 371–380, New York, NY, USA. Association for Computing Machinery.