DEBISS: a Corpus of Individual, Semi-structured and Spoken Debates

Klaywert Danillo Ferreira de Souza[1], **David Eduardo Pereira**[1],
Cláudio E. C. Campelo[1], **Larissa Lucena Vasconcelos**[2]

[1]Federal University of Campina Grande (UFCG) - System and Computing
Campina Grande, Brazil

[2]Federal Institute of Paraíba (IFPB)
Monteiro, Brazil

klaywertdanillo@copin.ufcg.edu.br , david.pereira@ccc.ufcg.edu.br

campelo@dsc.ufcg.edu.br, larissalucena@gmail.com

Abstract. *Debating is essential in daily life — whether in academic or professional settings, casual conversations, political forums, or online discussions. The range of debate applications is broad; therefore, their structures and formats can vary significantly. Developing corpora that account for these variations is challenging. The scarcity of debate corpora in the current state of the art, particularly for other languages beyond English, is notable. For this reason, this research proposes the DEBISS corpus, a collection of spoken and individual debates in Portuguese with semi-structured features. The corpus has broad applicability across Natural Language Processing tasks, including speech-to-text, speaker diarization, argument mining, and debate quality evaluation.*

1. Introduction

Debates have been an essential part of human communication and decision-making since ancient Greece [Bentahar et al. 2010], where philosophers such as Socrates, Plato, and Aristotle engaged in discussions that profoundly influenced Western thought. Over time, debates have become a fundamental aspect of daily life, offering numerous benefits such as enhancing communication skills, fostering critical thinking, and improving persuasive abilities.

Despite their importance, progress in the computational analysis of debate is constrained by a scarcity of annotated corpora. This limitation restricts research opportunities and the development of specialized computational models capable of processing nuanced debate interactions. Moreover, there is a notable lack of resources for oral and semi-structured debates, particularly in languages other than English. To address this gap, we introduce DEBISS (Spoken, Individual, and Semi-structured Debates), a novel annotated Brazilian Portuguese corpus. It contains **9 hours and 35 minutes** of debates on the theme ‘Generative Artificial Intelligence and its impacts on society”, carried out among **67** first-semester computer science students from the Federal University of Campina Grande.

The DEBISS corpus focuses on spoken, individual debates that combine predefined questions with opportunities for open expression, enabling spontaneous argumentation distinct from highly structured or written formats. This approach more closely reflects debates in educational and everyday contexts and differs significantly from political debates or discussions on social networks. The key contributions of this corpus include its focus on Brazilian Portuguese, its debate format that extends beyond those explored in the state of the art, and its grounding in an educational setting, offering insights into student oratory skill development. The dataset is further enriched with detailed self and peer evaluations regarding performance and topic knowledge, along with multimodal data (audio and transcriptions) essential for comprehensive analysis.

Therefore, this comprehensive resource is valuable for several Natural Language Processing (NLP) tasks, including Argument Mining (AM), Debater Quality Analysis (DQA), speech-to-text, speaker diarization, disfluency detection, and more, making a significant contribution to NLP research in Brazilian Portuguese. The DEBISS corpus, including all transcriptions, audio files, and associated annotations, will be made available for research purposes via GitHub[1].

2. Related Work

Debates analysis has attracted attention across various fields, including computational linguistics, discourse analysis, and Artificial Intelligence (AI). Political debates have been a primary focus for corpus construction due to their structured nature and public availability. Notable examples include the U.S. Presidential Debate Corpus [Vrana and Schneider 2017] and other political datasets [Duthie et al. 2016, Mestre et al. 2021, Mancini et al. 2022], which comprise transcripts annotated for argumentative structures and rhetorical strategies [De Smedt and Jaki 2018, Carvalho et al. 2011, Hautli-Janisz et al. 2022]. Despite their utility, these corpora often exhibit high formality and strict protocols, which may not reflect the dynamics of spontaneous interactions. Their focus on specific structures and topics also limits model generalization to other debate types.

Furthermore, the popularization of social networks has led to development of corpora from online discussions and debates [Durmus and Cardie 2019, Khodak et al. 2017, Stranisci et al. 2021, Lai et al. 2018], often focusing on written exchanges on platforms like Twitter and Reddit [Sousa et al. 2021, Habernal and Gurevych 2016, Boltužić and Šnajder 2016, Chakrabarty et al. 2019]. The Internet Argument Corpus (IAC) [Abbott et al. 2016] aggregates discussions annotated for stance, quality, and relevance, covering diverse topics and styles. While rich for studying informal discourse and sentiment analysis, these datasets suffer from noise, informal language, and inconsistent structures, and they lack multimodal and spontaneous elements of spoken debates.

Academic settings offer ground for studying the development of argumentative skills in students. Most educational corpora [Ruiz-Dolz et al. 2021] include recordings and transcripts of formal debates, offering insights into educational discourse but rarely addressing less structured formats. There is a scarcity of annotated datasets capturing nuances of student-led debates, particularly in languages other than English, which limits cross-cultural and multilingual research in argumentative discourse analysis.

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

3. Methodology

The DEBISS corpus comprises audio transcriptions of in-person debates conducted with the consent of first-year computer science undergraduate students at Federal University of *Campina Grande*. Each debate session was moderated by a facilitator. Data collection took place in 2024 and involved **67** students, who were organized into **16** debate groups, generating a total of **9 hours** and **35 minutes** of audio recordings. These recordings were transcribed using a semi-automated process that combined speech-to-text AI models with human validation to ensure accuracy. In this study, debate groups consisted of **3** to **5** participants, totaling 16 groups. Each participant defended their own viewpoint, speaking independently. Table 1 provides a summary of the corpus stats following the application of the complete methodology.

Table 1. DEBISS stats numbers

Metric	Value
Number of debaters	67
Number of groups	16
Total audio length of recordings	9 hours and 37 minutes
Number of tokens	130697
Words lexical diversity[2]	0.062

Debate Subject To create a focused and engaging debate, a central theme was selected: “Generative Artificial Intelligence and Its Impacts on Society”. This topic is highly relevant and controversial, prompting extensive questioning, criticism, across various fields. Furthermore, to stimulate the debate, a collection of online texts were compiled to be shared with the debaters ahead of the debate. These were selected for their readability, prioritizing news articles and opinion pieces from reputable sources over complex academic papers. The texts addressed specific, relevant, and controversial issues related to the theme, such as the impact of AI and the its legal implications. Since reading the texts was voluntary, it was also provided a two-page summary which condensed key information on critical topics.

Debate Environment Setup To collect the data, the debate sessions were recorded using a Logitech USB Yeti Condenser Microphone in omnidirectional mode, paired with OBS Studio software. The audio was captured in a 3x5m conference room where debaters sat around a central table. OBS was configured to the optimal decibel level for all participants, producing MP3 audio files. The room also included a 55-inch TV displaying information that was used to guide the debate section.

Debate Format The data collection protocol was designed to capture speech data. Each session begins with an explanation of the study’s purpose, informing the voluntary participants that the debate will be recorded without including any personal information. The

[2] Measure of vocabulary variation within a text

primary goal is to create a transcribed debate corpus for use in various scenarios, particularly in developing AI models focused on debate analysis.

Next, participants are then asked to sign a consent form authorizing the use of their recorded voices for research purposes. Additionally, we took care to anonymize the data. The debaters were identified only by numerical identifiers and not by their personal names. Each debater initially records the same sentence before the debate begins, for identification purposes. Thereafter, a moderator oversees the debate, ensuring order and encouraging discussion. The moderator first explains the rules: debaters should listen to the moderator, refrain from interrupting each other, and raise their hand if they wish to speak, waiting for their turn or for the current speaker to finish.

Also, the debate section is divided into three parts. The **first part** allows debaters to express their initial opinions on the topic. The **second part** consists of a question-and-answer round, and the **third part** involves final thoughts and reflections on the topic. Each debater is assigned specific questions, which are read aloud by the moderator and displayed on a TV. After answering, other debaters may comment or ask additional questions. This interaction is optional, providing an open space for engagement while maintaining a semi-structured format with both mandatory and voluntary interactions. Each debater has uninterrupted time to express their opinions and answer questions. A total of five questions are prepared, with one question per debater; if there are only three participants, only the first three questions are used.

Afterward, there is a final question for all participants, allowing those who wish to contribute with additional insights. This interaction is optional, fostering participatory debate that encourages the exchange of diverse ideas. In the final part, participants are asked to share any concluding thoughts and whether their opinions have changed based on the discussion. This final reflection is mandatory, giving each debater an opportunity to express their final stance.

Debater Evaluation Once the debate and recording ended, each participant completed a self-evaluation form (Google Forms). This was crucial for assessing individual performance and personal reflections on the debate process, providing valuable data for analysis. The self-evaluation form included the following questions:

- *What is your identification number?* (Multiple choice)
- *How would you rate your performance during the debate?* (Likert Scale: Poor to Excellent)
- *Before the debate, how well did you know the topic (from study or experience)?* (Likert Scale: Strongly Disagree to Strongly Agree)
- *Did you thoroughly read the material provided by the organizers?* (Likert Scale: Strongly Disagree to Strongly Agree)
- *Did you study additional materials beyond what was provided?* (Likert Scale: Strongly Disagree to Strongly Agree)
- *Did the debate expand your perspective or knowledge on the topic?* (Likert Scale: Strongly Disagree to Strongly Agree)
- *Who do you think was the best debater?* (Multiple Choice, including the option ”there was no best debater”)
- *Justify your choice for the best debater.* (Open Text)

Participants rated their own performance, reflecting on their effectiveness in argumentation, which provided insights into their self-perception within the debate dynamics. Questions about prior knowledge and preparation assessed the participants' familiarity with the topic and their engagement with the provided materials. This helped to assess how preparation influenced their performance. The question on whether the debate broadened their perspective aimed to measure the educational value of the session.

Furthermore, the debater evaluation included a peer assessment: participants identified the best debater among the group and justified their choice, this data is valuable for analyzing the results of debates and understanding the factors that contribute to successful debates. Also, it is crucial for participants' development, fostering meta cognition and self-improvement.

Data Processing and Annotation After recording all debate sessions, both automatic and manual post-processing were necessary. The first step was transcribing the recorded audio files, a time-consuming task. To expedite this, we used automatic transcription with free models, testing three speech to text models: wav2vec-large [3], whisper-large [4], and Azure Speech-to-Text [5]. After comparing their performance on the same recordings, we found Azure's model to be the most accurate and suitable for our context. All audio files were then transcribed using Azure, with the outputs exported to CSV files containing the text transcribed from the audio chunk, the start time of the transcribed audio chunk and the end time of the transcribed audio chunk.

Errors were manually corrected in the transcriptions. Google Sheets was used as the annotation tool for editing of text and timestamps. Annotators followed a specific protocol to correct errors in the transcriptions. The primary goal was to ensure accuracy by listening to the audio, comparing it to the automatic transcription, making necessary corrections. Moreover, the transcriptions were kept literal to preserve a faithful representation of spoken data, including speech disfluencies.

Also, a final annotation was done to identify the speakers in the audio. A new column was added to the spreadsheet to label each piece of speech with the corresponding debater's identifier. Additionally, if the automatic transcription split speech into smaller fragments due to pauses, annotators combined these fragments into a single line, adjusting the start and end times accordingly. Figure 2 illustrates the entire annotation process for this dataset.

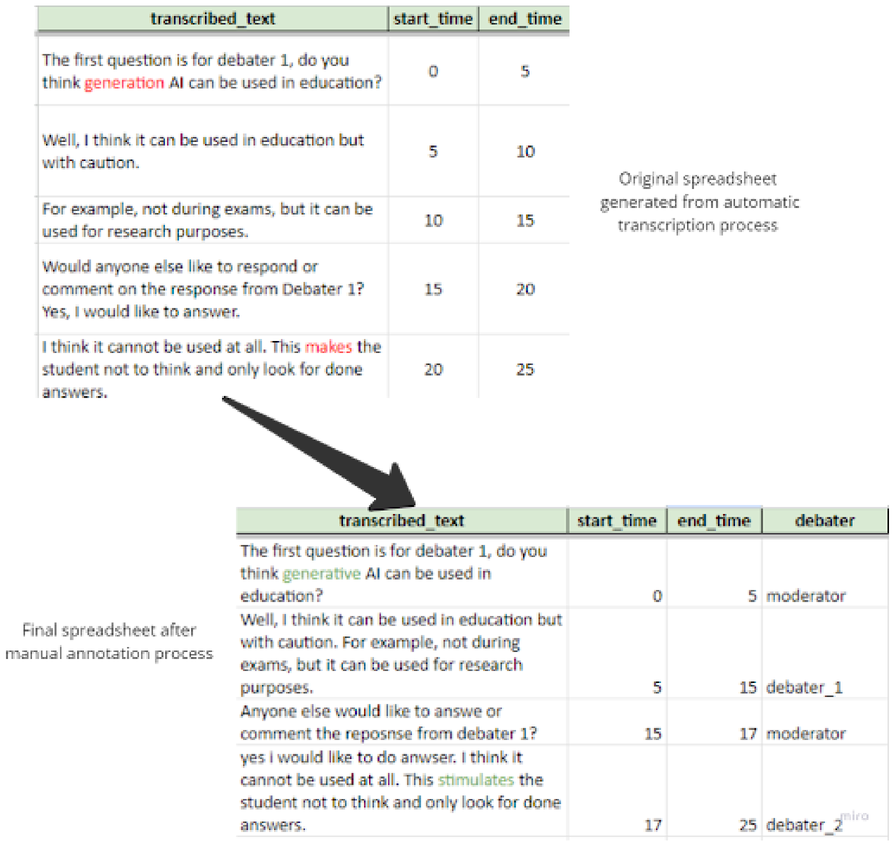
4. Conclusion

The proposed corpus was developed to advance the state of the art by introducing a new audio-recorded resource with strong potential to enhance the scientific study of debate analysis. DEBISS distinguishes itself from existing debate corpora through features such as semi-structured formats, individual debates, and spoken content collected in educational contexts. Beyond the dataset itself, this work also contributes a clear methodology for gathering debates in this specific format.

[3] https://huggingface.co/facebook/wav2vec2-large

[4] https://huggingface.co/openai/whisper-large-v3

[5] https://learn.microsoft.com/en-us/azure/ai-services/



transcribed_text	start_time	end_time
The first question is for debater 1, do you think generation AI can be used in education?	0	5
Well, I think it can be used in education but with caution.	5	10
For example, not during exams, but it can be used for research purposes.	10	15
Would anyone else like to respond or comment on the response from Debater 1? Yes, I would like to answer.	15	20
I think it cannot be used at all. This makes the student not to think and only look for done answers.	20	25

Original spreadsheet generated from automatic transcription process

transcribed_text	start_time	end_time	debater
The first question is for debater 1, do you think generative AI can be used in education?	0	5	moderator
Well, I think it can be used in education but with caution. For example, not during exams, but it can be used for research purposes.	5	15	debater_1
Anyone else would like to answe or comment the reposnse from debater 1?	15	17	moderator
yes i would like to do anwser. I think it cannot be used at all. This stimulates the student not to think and only look for done answers.	17	25	debater_2

Final spreadsheet after manual annotation process

Figure 1. Annotation process example

To highlight the corpus relevance, it important to mention DEBISS's applicability in related studies with specific objectives. Based on the original DEBISS corpus, it was developed DEBISS-Arg [Pereira et al. 2025], a fully annotated dataset for AM with detailed labels for Argument Discourse Units, argumentative components, and micro/macro-level relations between debaters' statements. Furthermore, introduced DEBISS-Eval [6], a subcorpus designed for debate quality assessment through expert evaluations, offering both quantitative scores and rich qualitative feedback on debaters' skills. Additionally, DEBISS has been applied in text disfluency detection [Lima and Campelo 2024], where advanced LLMs — particularly GPT-4o — showed strong performance in identifying and removing disfluent elements from debate transcripts. Moreover, the corpus includes rich annotations supporting multiple NLP tasks, including speech-to-text transcription, voiceprint identification, speaker diarization, and silence detection.

Although the dataset follows a well-defined methodology and presents comprehensive statistical data, some limitations should be acknowledged. One notable constraint is the narrow thematic scope of the debates, which may limit the dataset's generalizability. Expanding the range and diversity of topics would improve its applicability. Moreover, collecting data from more heterogeneous participant groups—such as students from varied educational backgrounds and levels—would enhance the dataset's representativeness. These limitations can be addressed in future work by applying the proposed methodology to a broader set of themes and more diverse debater profiles.

[6] https://github.com/AINDA-Project-UFCG/debater-quality-data

References

[Abbott et al. 2016] Abbott, R., Ecker, B., Anand, P., and Walker, M. (2016). Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452.

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

[Boltužić and Šnajder 2016] 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.

[Carvalho et al. 2011] Carvalho, P., Sarmento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 564–568.

[Chakrabarty et al. 2019] 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.

[De Smedt and Jaki 2018] De Smedt, T. and Jaki, S. (2018). The polly corpus: Online political debate in germany. In *of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)*, page 33.

[Durmus and Cardie 2019] Durmus, E. and Cardie, C. (2019). A corpus for modeling user and language effects in argumentation on online debating. *arXiv preprint arXiv:1906.11310*.

[Duthie et al. 2016] 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.

[Habernal and Gurevych 2016] 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.

[Hautli-Janisz et al. 2022] 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).

[Khodak et al. 2017] Khodak, M., Saunshi, N., and Vodrahalli, K. (2017). A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*.

[Lai et al. 2018] Lai, M., Patti, V., Ruffo, G., and Rosso, P. (2018). Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 15–27. Springer.

[Lima and Campelo 2024] 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.

[Mancini et al. 2022] 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 et al. 2021] 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.

[Pereira et al. 2025] Pereira, D., Simão, D., and Claúdio, C. (2025). Debiss-arg: An in depth data annotation protocol and corpus for argument mining in semi structured debates. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC.

[Ruiz-Dolz et al. 2021] Ruiz-Dolz, R., Nofre, M., Taulé, M., Heras, S., and García-Fornes, A. (2021). Vivesdebate: A new annotated multilingual corpus of argumentation in a debate tournament. *Applied Sciences*, 11(15):7160.

[Sousa et al. 2021] 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.

[Stranisci et al. 2021] Stranisci, M., De Leonardis, M., Bosco, C., and Patti, V. (2021). The expression of moral values in the twitter debate: a corpus of conversations. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):113–132.

[Vrana and Schneider 2017] Vrana, L. and Schneider, G. (2017). Saying whatever it takes: Creating and analyzing corpora from us presidential debate transcripts.