ARGUMENT-BASED DETECTION AND CLASSIFICATION OF FALLACIES IN POLITICAL DEBATES

Pierpaolo Goffredo[1], Mariana Chaves Espinoza[1], Elena Cabrio[1], Serena Villata[1]
[1]Université Côte d'Azur, CNRS, Inria, I3S, France

INTRODUCTION

Fallacious arguments are often used in political debates to mislead, detecting and classifying **fallacies** is an important open challenge in **Argument Mining** (AM) and **NLP** research to limit their potential harmful **influence**.

Our contribution is twofold:

- **Extension** of the dataset ElecDeb60to16
- Transformer-based **model for detecting and classifying fallacies** in political debates

Our results show the advantages of complementing transformer-generated text representations with **non-textual features**.

ELECDEB60TO20

To effectively address the task of detecting and classifying fallacious arguments within political debates, we decided to rely on the **ElecDeb60To16** dataset.
We expanded the dataset with the **transcripts** of the debates of this election campaign to include updated annotations, incorporating argumentative **components**, as well as the **relations** between these components, and **fallacies**.

As a result of this annotation update, the dataset is renamed as **ElecDeb60to20**, reflecting the coverage of debates spanning from 1960 to 2020.

ANNOTATION

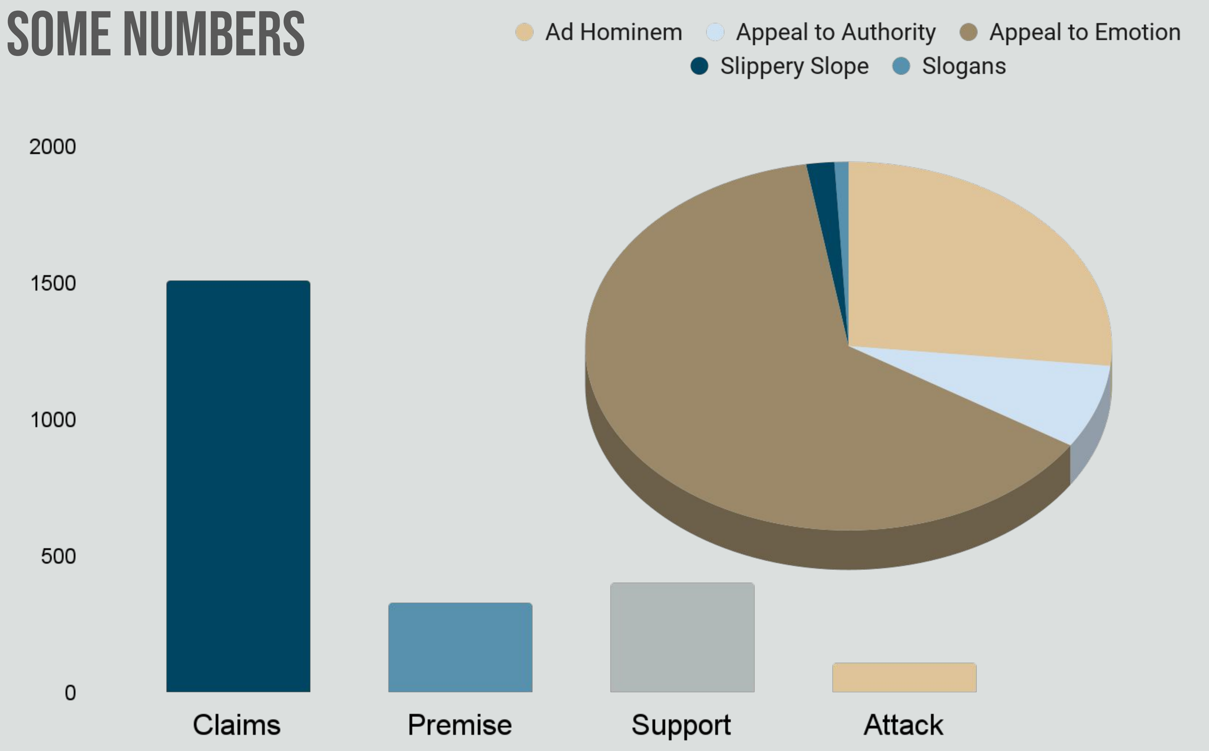
We annotated **Trump vs. Biden** 2020's debate considering:

- **Fallacy** (six categories)
- Argumentative **component**
- Argumentative **relations**

A set of 50 sentences randomly extracted from the debates was annotated to assess **Inter-Annotator Agreement** (IAA):

- Observed Agreement 0.857
- Krippendorff's α 0.757

SOME NUMBERS



METHOD & RESULTS

We cast the fallacy detection task as an **Information Extraction** problem, where the goal is to identify and classify in the debates the **textual snippets** corresponding to the six categories of fallacies annotated in the context of a political debate.
We employ transformer-based architectures in both their *basic* configuration and in a *specialized* configuration designed for token classification.

We build a **contextual framework** that includes the sentence containing the fallacy, as well as the preceding and following sentences. When the fallacious sentence is the first or last in the dialogue, the preceding or following sentence is excluded.

Moreover, we **enhance** the specialized architecture by including **non-textual features**: argumentative components, argumentative relationship, and Part of Speech tags.

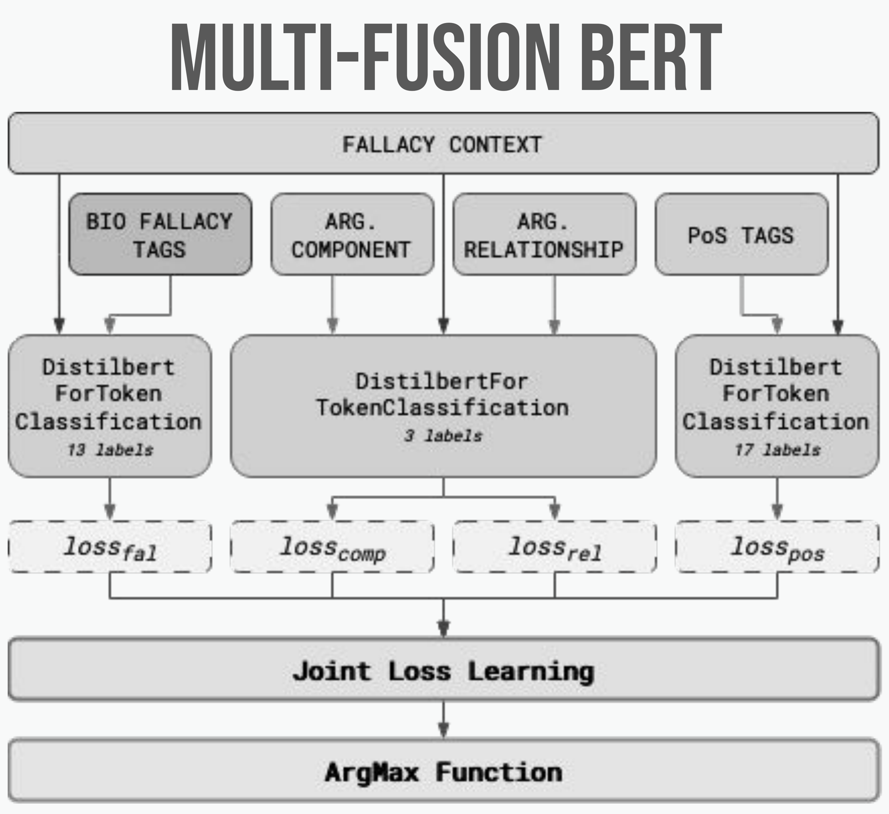
All the features are combined in order to calculate a **joint loss** as such: $joint_{loss} = \alpha * \frac{(loss_{fal} + loss_{cmp} + loss_{rel} + loss_{PoS})}{N_{loss}}$

Multi-Fusion BERT computes logits (L) for each feature by employing a specialized **TokenForClassification** Transformer model adapted to the number of labels:

- 3 for components
- 3 for relations
- 17 for PoS

The **architectures** for argumentative features for components and relations share the same parameters, enabling us to obtain **logits** for both components and relations. An additional model, based on the number of PoS tags (i.e., 17), is used to obtain logits for PoS features.

Distinct **losses** are computed for each model: *fallacy loss, component loss, relation loss, and part-of-speech loss.*



MODEL	AVG MACRO F1 SCORE
BERT + LSTM	0.4697
BERT + LSTM (comp. and rel. feat.)	0.5142
BERT + BiLSTM + LSTM	0.5495
BERT + BiLSTM + LSTM (comp. and rel. feat.)	0.5614
BERT FTC (bert-base-uncased)	0.7096
BERT FTC (bert-large-cased-finetuned-conll03-english)	*0.7237*
DeBERTa FTC (microsoft/deberta-base)	0.7222
Electra FTC (electra-base-discriminator-finetuned-conll03-english)	0.4033
DistilBERT FTC (distilbert-base-cased)	0.7010
DistilBERT FTC (distilbert-base-uncased)	0.7047
MultiFusion BERT (comp. & rel. & PoS)	**0.7394**

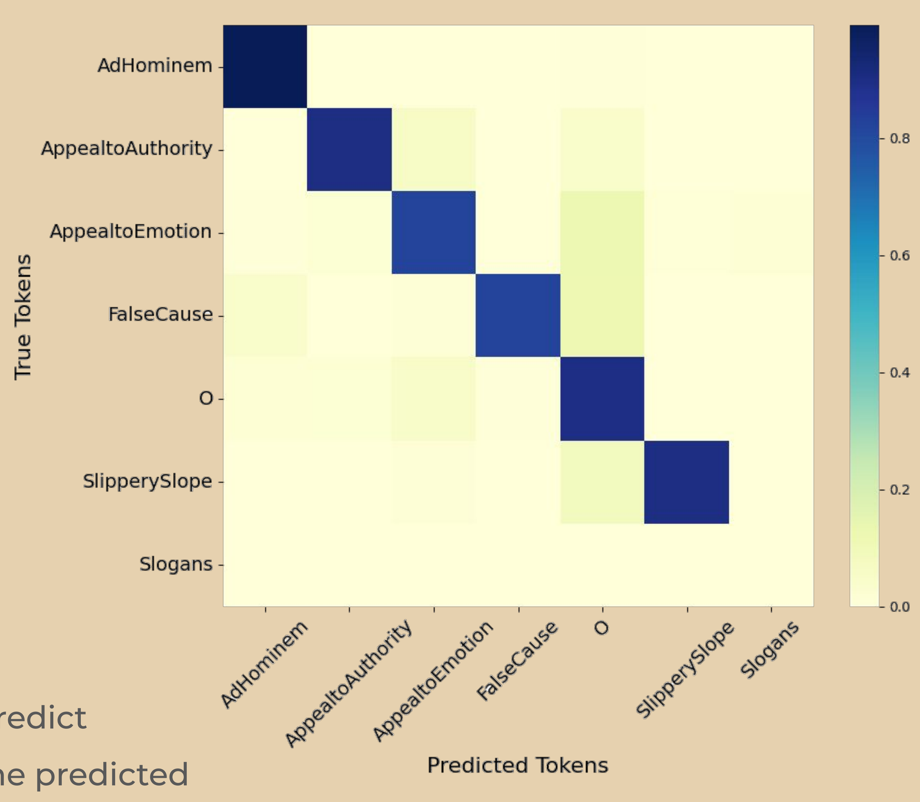
COMPONENT	RELATIONSHIP	PART OF SPEECH	AVG MACRO F1 SCORE
✅			0.6922
	✅		0.6922
		✅	0.7212
✅	✅		0.7278
✅		✅	0.7166
	✅	✅	0.7166
✅	✅	✅	**0.7394**

EVALUATION

Despite the relatively smaller size of the dataset and the task complexity, the **results** obtained from the different models are **promising**. To better analyze the impact of the different features incorporated in our architecture, we carried out **ablation tests**. The table above presents the results obtained by **MultiFusion BERT** using all possible combinations of:

- Argumentative components
- Argumentative relations
- Context PoS tags

Notably, the identification of tokens labeled as ***Slogans*** exhibits the poorest results, despite being relatively easier to recognize for humans. This can be due to the limited presence of examples/tokens in both the training and the test set. On the contrary, tokens labeled as ***Slippery Slope*** and ***False Cause*** are much better classified.



The confusion matrix reveals that the model tends to over-predict instances in ***Other*** category. As observed in the column of the predicted O class, **false positives** are the most prevalent in the non-fallacious tokens. Moreover, ***False Cause*** and ***Appeal to Emotion*** are the classes that the models misinterpret the most as non fallacious. In a smaller proportion, the model misclassified instances of ***Appeal to Authority*** as ***Appeal to Emotion***.

The results obtained for the other labels are **in line** with those in (Goffredo et al., 2022) for the **classification** task only. The addition of new fallacious examples from the 2020 debates kept **unchanged** the distribution of fallacies with respect to the **previous** debates, suggesting that the detection of fallacious snippets remains **consistent** and **stable** across different debate **contexts**.

CONCLUSION & FUTURE WORK

This paper enhances existing argumentation schemes (Walton, 1995) for real-world scenarios, like political debates, by **extending** the ElecDeb60to16 dataset to include the Trump vs. Biden 2020 debate. It introduces **MultiFusion BERT**, a transformer-based model that **combines** debate **text** and **features** for efficient fallacy detection and classification.

In future research, we aim to explore more **complex** fallacy categories, like causal fallacies, by integrating **knowledge** and **reasoning features**. Our goal includes **generating** valid arguments from identified fallacies and addressing the formal invalidity of fallacious arguments through the **creation** of new, valid **arguments**.

REFERENCES

- Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. **Fallacious argument classification in political debates**. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Elena Cabrio and Serena Villata. 2018. **Five years of argument mining: a data-driven analysis**. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, pages 5427–5433. ijcai.org.
- Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. **Yes, we can! mining arguments in 50 years of US presidential campaign debates**. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.
- Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2021. **"Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification**. In RANLP 2021 - Recent Advances in Natural Language Processing, Varna / Virtual, Bulgaria.
- D.N. Walton. 1987. Informal Fallacies: **Towards a Theory of Argument Criticisms. Pragmatics & beyond companion series**. J. Benjamins Publishing Company.
- D.N. Walton. 1995. **A Pragmatic Theory of Fallacy. Studies in rhetoric and communication**. University of Alabama Press.