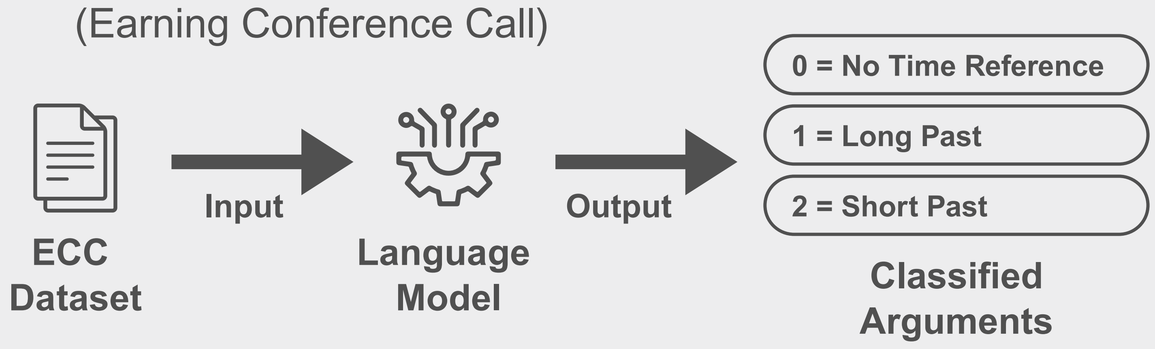
AIDAVANCE at the NTCIR-18 FinArg-2 Task: Making the Most of Small Language Models

Hugo Dutra (1), Leonardo Martinho (1), Gabriel Assis (1), Jonnathan Carvalho (2), Aline Paes (1)

(1) Instituto de Computação, Universidade Federal Fluminense (UFF) , Niterói, RJ, Brazil
(2) Instituto Federal Fluminense (IFF), Campus Itaperuna, RJ, Brazil
hugo_dutra@id.uff.br, leonardoalvesmartinho@id.uff.br, assisgabriel@id.uff.br, joncarv@iff.edu.br, alinepaes@id.uff.br

1. FINARG-2 TASK

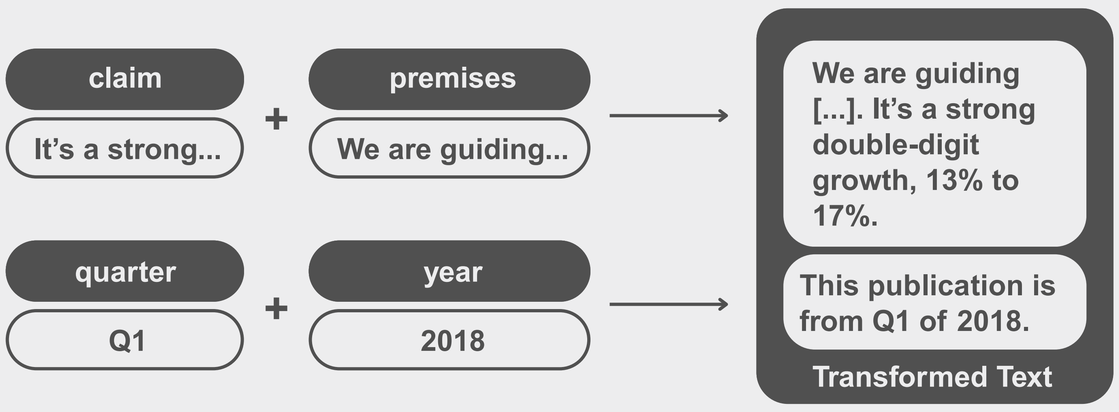
- Detection of Argument Temporal References (Earning Conference Call)



- We evaluated two primary strategies:
  - **Fine-tuning BERT Models**
  - **Prompt-Based Classification with LLMs**

2. DATA PREPROCESSING

- **"claim"** and **"premises"** were merged into full, cohesive paragraphs, based on gramatical cohesion;
- **"quarter"** and **"year"** were then appended to the end these paragraphs as **publication dates**.



3. FINE-TUNING BERT MODELS

3.1. Model Selection

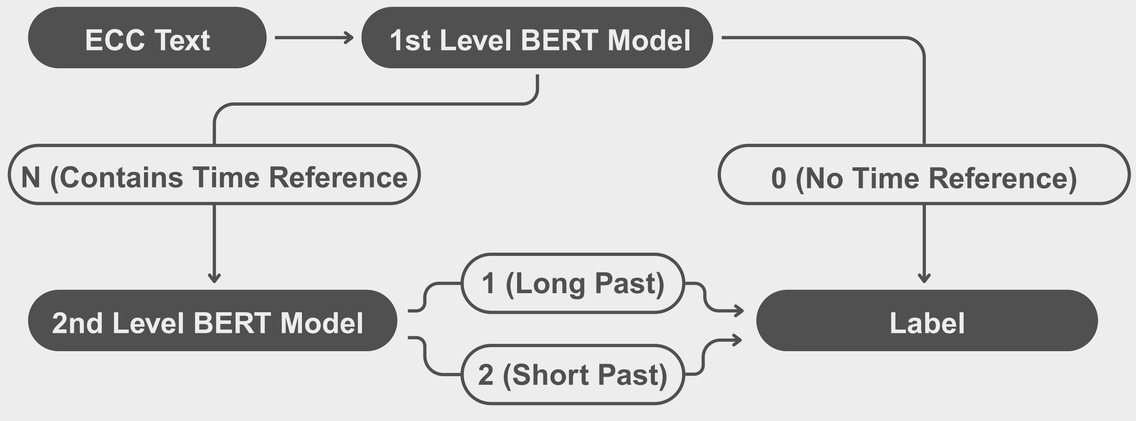
- Explored models include DeBERTa, mDeBERTa, DeBERTa-NLI, mDeBERTa-NLI and FinBERT.

3.2. 3-Label Classification Approach

- A single fine-tuned BERT-based model classifies texts into the original 3 categories.

3.3. Cascade Classification Approach

- A two-Step classification process with two fine-tuned BERT models.



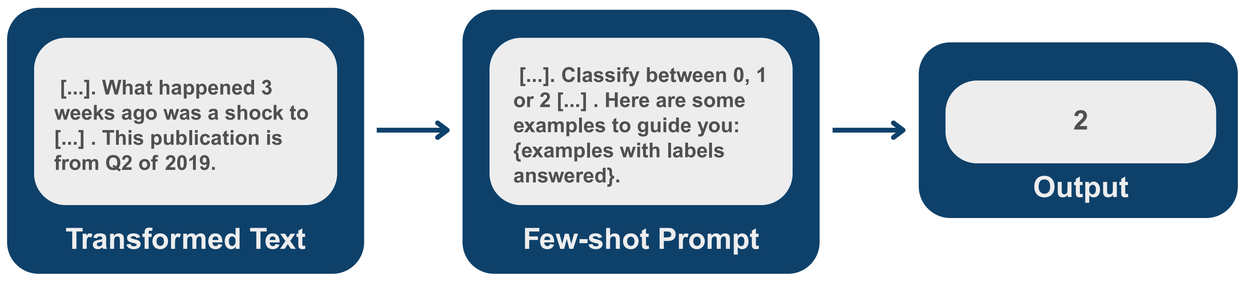
4. PROMPT-BASED LLM CLASSIFICATION

4.1. Model Selection

- LLMs used include OpenAI's (gpt-4o) and Google's (gemini-2.0-flash)

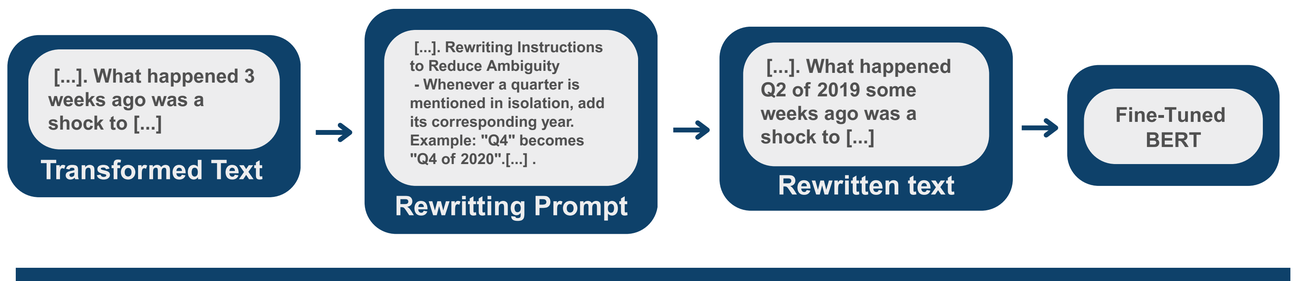
4.2. Few-shot Approach

- Few-shot text classification using Large Language Models.



4.3. Argument Rewriting Approach

- Rewriting ECC texts using LLMs, in order to remove time ambiguities and further enhance BERT models performance. This approach led to no significant classification improvements.



5. RESULTS

5.1. Experimental Results

- The following table portrays each model's validation results:

	3-Label Approach			Cascade Approach		
Model	**Micro-F1**	**Macro-F1**	**W-F1**	**Micro-F1**	**Macro-F1**	**W-F1**
DeBERTa	0.7600	0.7278	0.7648	0.7400	0.7187	0.7501
DeBERTa-NLI	0.7600	0.7278	0.7648	0.7667	0.7508	0.7740
mDeBERTa	**0.8000**	**0.7878**	**0.8025**	0.7600	0.7416	0.7672
mDeBERTa-NLI	**0.7933**	**0.7705**	**0.7959**	**0.7733**	**0.7593**	**0.7802**
Finbert	0.7333	0.7010	0.7391	N/A	N/A	N/A
GPT-4o (Few-shot)	0.6933	0.6725	0.7050	0.6733	0.6445	0.6841
GPT-4o (Base Model)	0.2533	0.2388	0.1889	0.5267	0.4140	0.5672

5.2. Official Submissions

- The **3 models** which achieved best overall results during validation were used as our task submissions. Their official scores and ranking can be observed in the table below:

Submissions	Micro-F1	Macro-F1	Ranking
AIDAVANCE_ECC_1 (3-Label mDeBERTa)	0.6905	0.6711	4th
AIDAVANCE_ECC_2 (3-Label mDeBERTa-NLI)	0.6667	0.6105	13th
AIDAVANCE_ECC_3 (Cascade mDeBERTa-NLI)	0.6905	0.6610	7th

6. CONCLUSIONS

- Fine-tuned **BERT models** consistently **outperformed** our **LLM-based** approaches
- **Multilingual** models such as mDeBERTa **outperformed monolingual** ones during validation
- **Finbert**, despite its **financial domain** specialization, consistently **underperformed** in this task

Acknowledgments

CAPES CNPq FAPERJ