SEMEVAL 2024 TASK 4 ON "MULTILINGUAL DETECTION OF PERSUASION TECHNIQUES IN MEMES”

EURECOM
Sophia Antipolis

Meriem ABIDI
Meriem DIMASSI

- SemEval 2024 is a significant event in the field of computational semantics.
  It's a workshop aiming at exploring and evaluating the computational understanding of meaning in natural language.

- Memes are one of the most popular type of content used in online disinformation campaigns that are becoming more spread. They may influence many users through rhetorical and psychological techniques (Smears, Name calling …)

- The goal of this semester project was to build models for identifying these techniques:
  - In the textual content of a meme only.
  - In a multimodal setting in which both the textual and the visual content are to be analyzed together.

- In our work, we focused only on the textual content of memes.

RELATED WORK

A Survey On Computational Propaganda Detection

The paper focuses on Computational Propaganda which use automated or algorithm-driven techniques to manipulate and influence public opinion.

Used Models

- Binary classification using **TSHP-17** and **QPop corpora:**
  - logistic regression & SVM

- Classification using **PTC corporate**:
  - Binary classification & multi-label multi-class classifications & span detection task

⇒ BERT- based contextual gave the best-performance

SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images

The task focuses on:

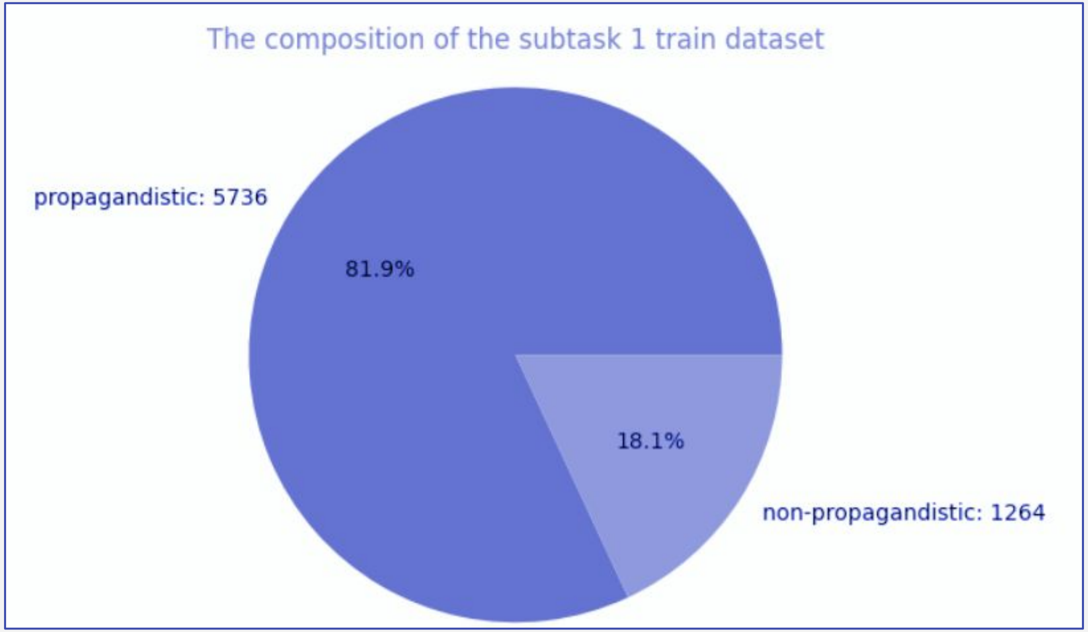
I. detecting the techniques in the text
II. detecting the text spans where the techniques are used
III. detecting techniques in the entire meme, i.e., both in the text and in the image.

Models per subtask

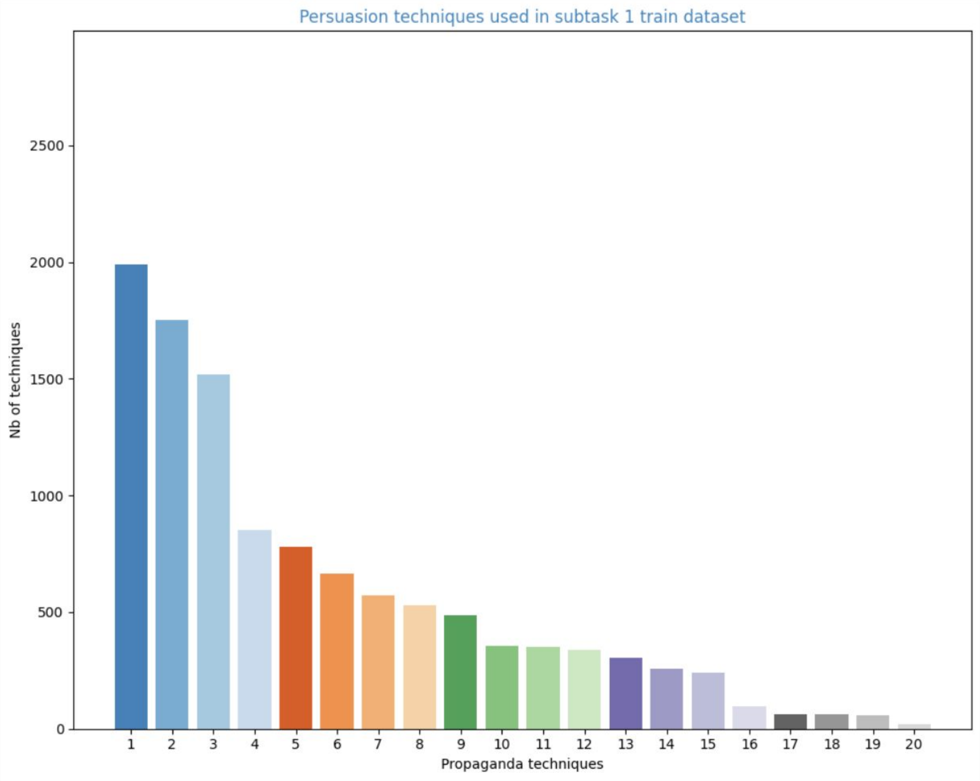
- **Subtask I** : most commonly used were RoBERTa, followed by BERT. Some participants used learning models such as LSTM, CNN, and CRF in their final systems
- **Subtask II**: BERT dominated, while RoBERTa was much less popular.
- **Subtask III (Multimodal: Memes)**: BERT dominated, but RoBERTa was quite popular as well.

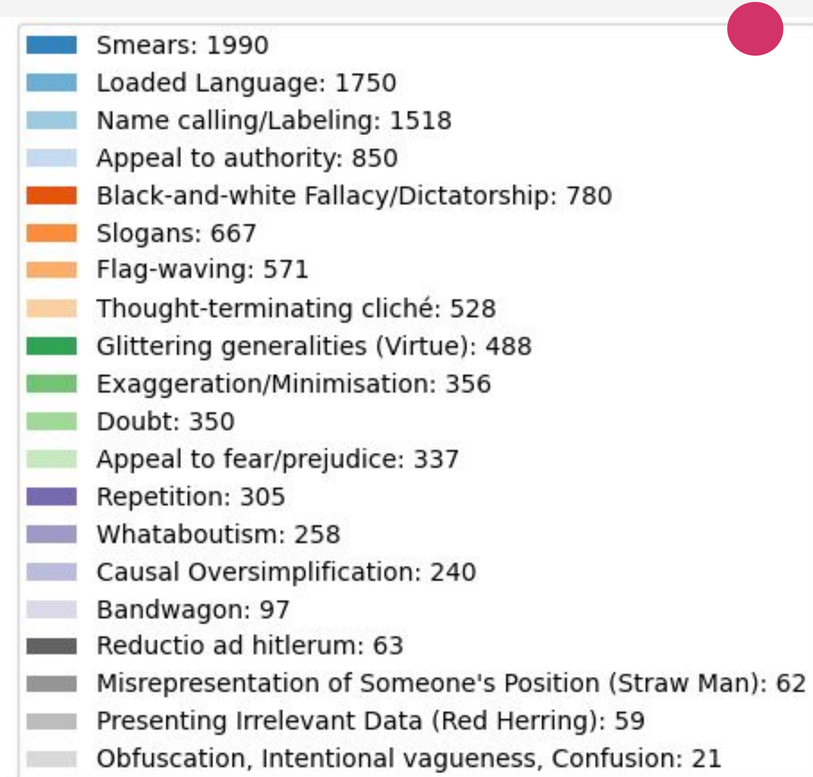
DATA ANALYSIS

Subtask1 - Train Dataset



Subtask1 - Train Dataset





🟦 Smears: 1990

🟦 Loaded Language: 1750

🟦 Name calling/Labeling: 1518

🟦 Appeal to authority: 850

🟧 Black-and-white Fallacy/Dictatorship: 780

🟧 Slogans: 667

🟧 Flag-waving: 571

🟧 Thought-terminating cliché: 528

🟩 Glittering generalities (Virtue): 488

🟩 Exaggeration/Minimisation: 356

🟩 Doubt: 350

🟩 Appeal to fear/prejudice: 337

🟪 Repetition: 305

🟪 Whataboutism: 258

🟪 Causal Oversimplification: 240

🟪 Bandwagon: 97

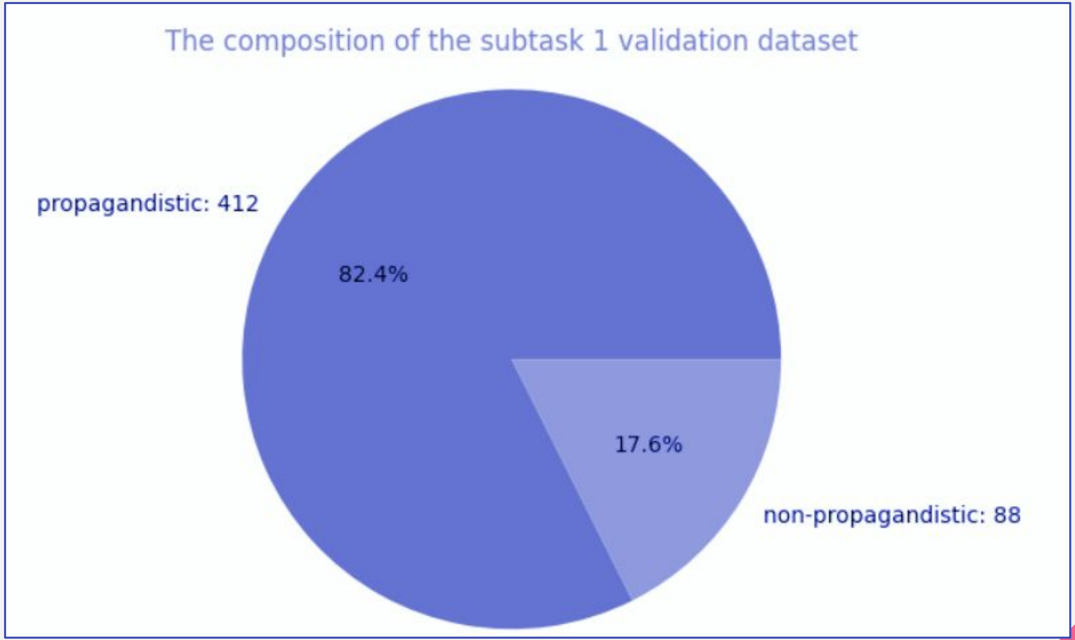
⬛ Reductio ad hitlerum: 63

⬛ Misrepresentation of Someone's Position (Straw Man): 62

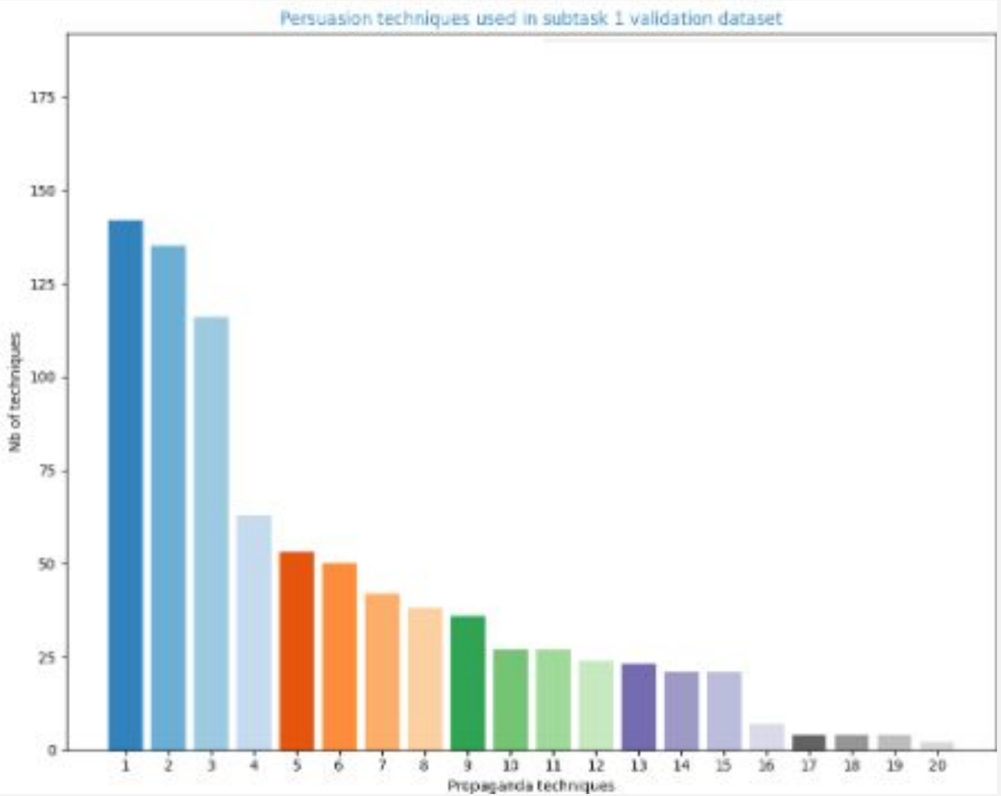
⬛ Presenting Irrelevant Data (Red Herring): 59

⬜ Obfuscation, Intentional vagueness, Confusion: 21

Subtask1 - Validation Dataset



Subtask1 - Validation Dataset



- 🟦 Smears: 142
- 🟦 Loaded Language: 135
- 🟦 Name calling/Labeling: 116
- 🟦 Appeal to authority: 63
- 🟧 Black-and-white Fallacy/Dictatorship: 53
- 🟧 Slogans: 50
- 🟧 Flag-waving: 42
- 🟧 Thought-terminating cliché: 38
- 🟩 Glittering generalities (Virtue): 36
- 🟩 Appeal to fear/prejudice: 27
- 🟩 Exaggeration/Minimisation: 27
- 🟩 Doubt: 24
- 🟪 Repetition: 23
- 🟪 Whataboutism: 21
- 🟪 Causal Oversimplification: 21
- 🟪 Bandwagon: 7
- ⬛ Misrepresentation of Someone's Position (Straw Man): 4
- ⬛ Reductio ad hitlerum: 4
- ⬛ Presenting Irrelevant Data (Red Herring): 4
- ⬛ Obfuscation, Intentional vagueness, Confusion: 2

PROPOSED APPROACH

BERT Model

- We used the **BertForSequenceClassification** from the Hugging Face transformers library, specifying the number of classes corresponding to our encoded labels.

- Each propaganda technique is represented by two classes, the first representing the "non-existence" of the technique and the second the "existence" of the technique.

- The model was refined using an **AdamW optimizer** and the **CrossEntropyLoss** function.

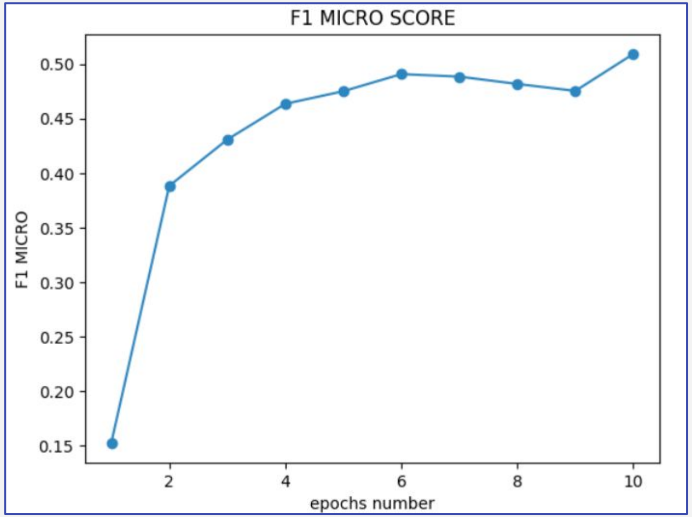
OPTIMIZATION OF MODEL PARAMETERS

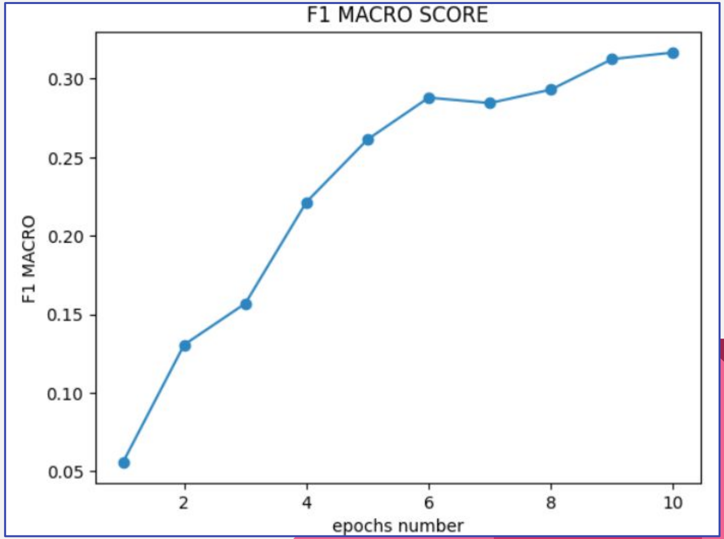
Number Of Training Epochs

- Insufficient number of training epochs ⇒ under-fitting
- Too many training epochs ⇒ over-fitting

Approach

We have tested the evolution of our model over a significant number of epochs in order to find a balance where the model achieves better results. We displayed **F1 Micro** and **F1 Macro** scores (Of BERT model with **lr=2e-5**) as a function of number of epochs, from **1 to 10.**





Classes Weights

Problem

- Variation in the frequency of different persuasion techniques in the dataset
⇒ Class imbalance ⇒ Decrease in the performance of the model.

Approach

- Assign weights to the classes that were inversely proportional to their frequency.
⇒ Give more weight to less frequent classes when calculating the loss ⇒ Improving the model’s ability to detect them.

Learning Rate

The learning rate determines the step size at each iteration while moving towards a minimum of a loss function.

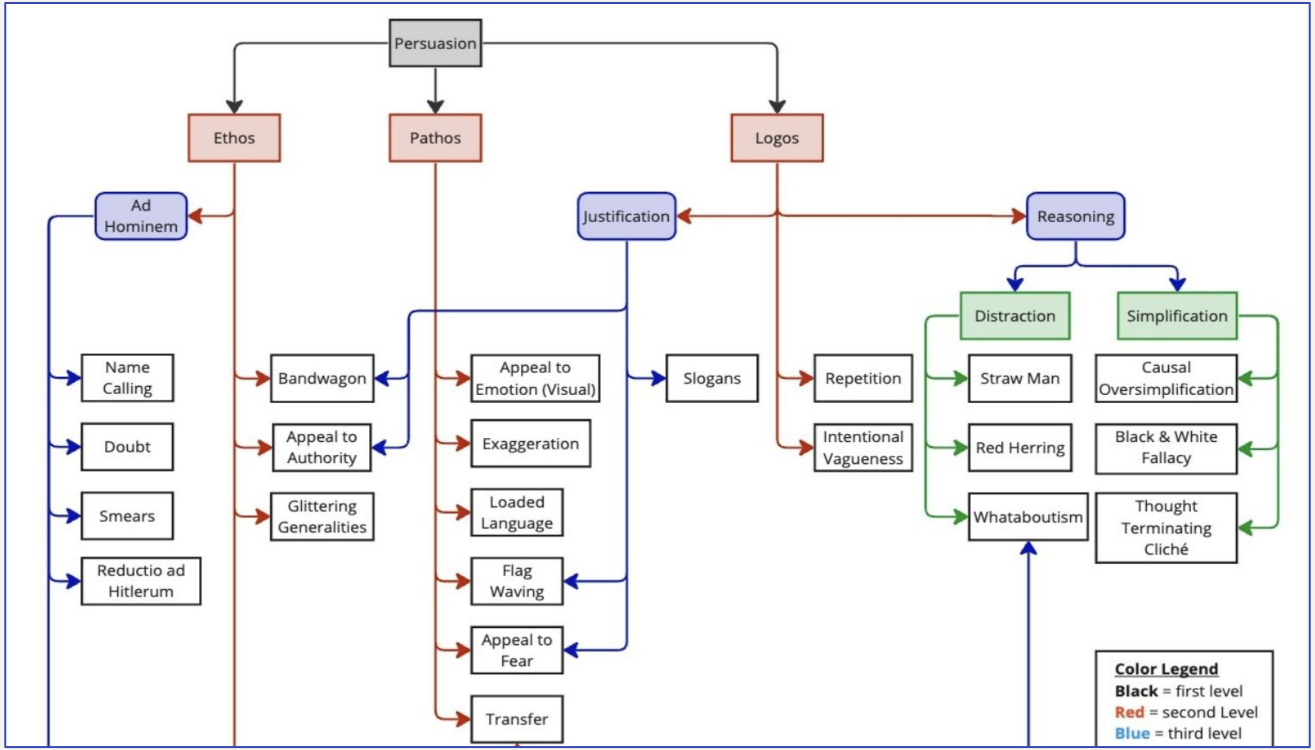
Problem

- Too high rate ⇒ The model converges too quickly to a suboptimal solution.
- Too low rate ⇒ The convergence time can be too long or the model can get stuck in local minima.

Approach

- For our BERT model, we tried different learning rates in order to optimise the training process.

DATA HIERARCHICAL ADAPTATION



- The hierarchy is essentially a directed acyclic graph that groups subsets of techniques sharing similar characteristics into a hierarchical structure.

- Based on this hierarchy, “Hierarchical F1” score is calculated to do the final ranking.

Our Hierarchical Approach

- Design the model’s training to respect the parent-child relationships between techniques.

A child not predicted + one of its parents detected ⇒ partial reward in the hierarchical F1 score

- Use **28** techniques Instead of working on the **20** techniques in text

⇒ Improve the model's ability to first identify larger categories of persuasion methods before focusing on more specific techniques.

COMBINING MULTIPLE APPROACHES

Combining BERT Without Weight Balancing And The Hierarchical Approach

- Classical BERT without weight balancing
  ⇒ model to prioritize the most frequent techniques and improve their detection.

- Hierarchical approach with weight balancing
  ⇒ Focuses on less common techniques while adding the ancestor nodes to the list.

How to combine these models?

compare the F1 score per propaganda technique at the end of the training phase for both models :

- ➢ difference >= 0.02 ⇒ we consider best performing model as the predictor for the class.
- ➢ difference < 0.02 ⇒ we consider the prediction of both models.

Combine BERT With Weight Balancing And BERT Without Weight Balancing While Parents Exist

- We implemented "parents detection model": a classical BERT model that only detects ancestor nodes.
  - These nodes are much more frequent than the other persuasion techniques ⇒ the model should be able to detect these techniques more accurately.

- A technique is confirmed as predicted only if:
  - It is identified by the most accurate model for that specific technique
  - At the same time, one of its "parent" techniques is recognized by the “parents detection model”.

RESULTS

Impact of Classes Weights

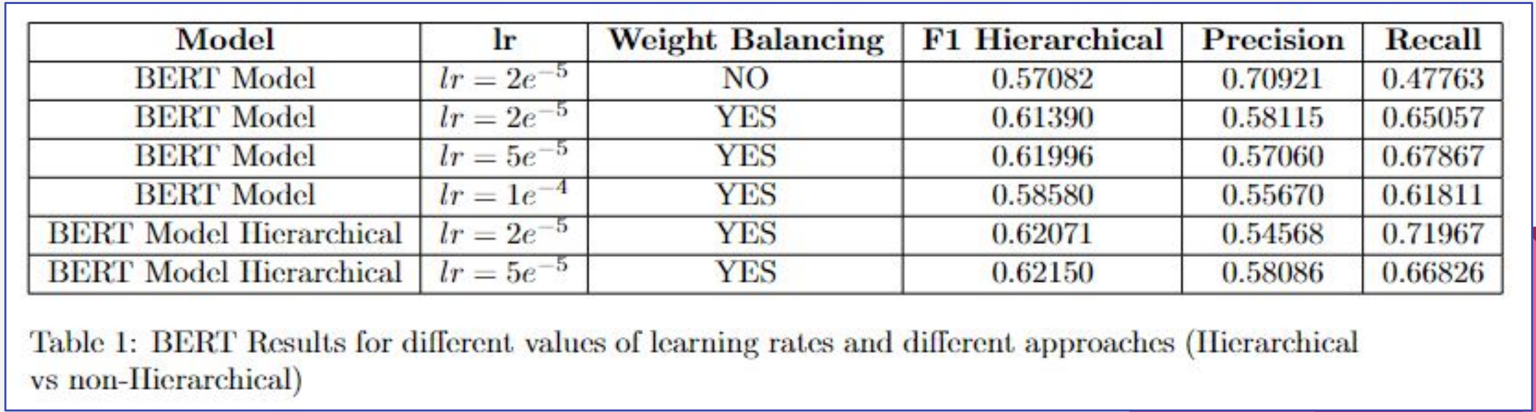
Results of Classical BERT when

- **There is no Weight Balancing** → F1 Hierarchical = 0.57082
- **We use weight balancing:** → F1 Hierarchical = 0.61390

Impact of the Learning Rate (lr)

We changed the learning rate value and we obtained the metrics for the BERT model using weight balancing for 3 different values:

- **lr = 5E-5** → F1 Hierarchical = 0.61996
- **lr = 2E-5** → F1 Hierarchical = 0.61390
- **lr = 1E-4** → F1 Hierarchical = 0.58580

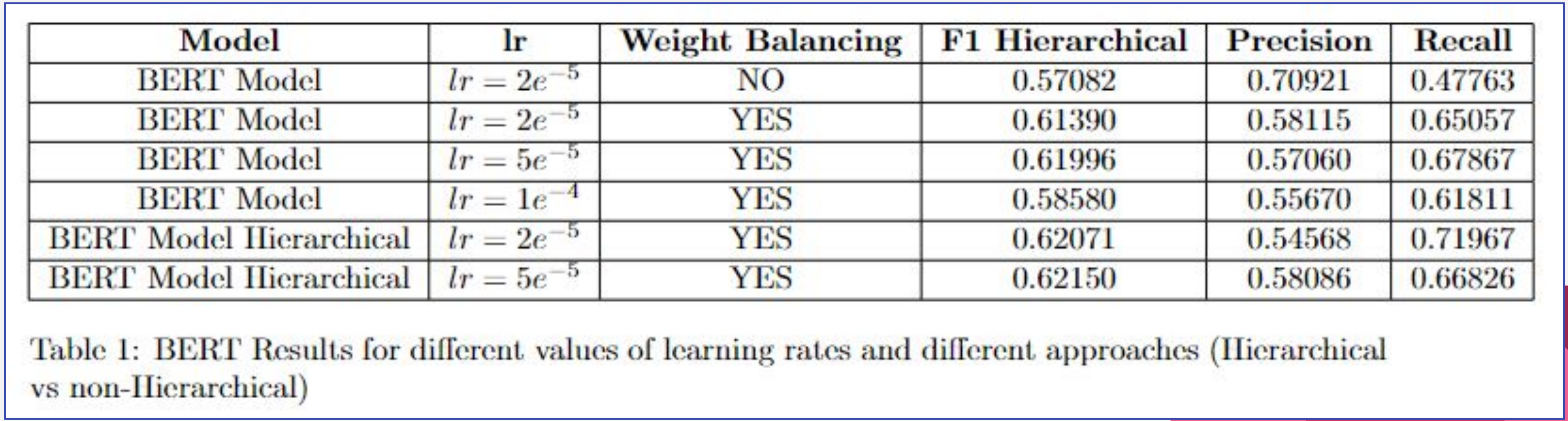


Model	lr	Weight Balancing	F1 Hierarchical	Precision	Recall
BERT Model	$lr = 2e^{-5}$	NO	0.57082	0.70921	0.47763
BERT Model	$lr = 2e^{-5}$	YES	0.61390	0.58115	0.65057
BERT Model	$lr = 5e^{-5}$	YES	0.61996	0.57060	0.67867
BERT Model	$lr = 1e^{-4}$	YES	0.58580	0.55670	0.61811
BERT Model Hierarchical	$lr = 2e^{-5}$	YES	0.62071	0.54568	0.71967
BERT Model Hierarchical	$lr = 5e^{-5}$	YES	0.62150	0.58086	0.66826

Table 1: BERT Results for different values of learning rates and different approaches (Hierarchical vs non-Hierarchical)

Our Hierarchical Approach

- For Learning Rate = 2e-5
  - **Classical BERT** → F1 Hierarchical = 0.61390
  - **Hierarchical BERT** → F1 Hierarchical = 0.62071
- For Hierarchical Approach
  - **lr = 2e-5** → F1 Hierarchical = 0.62071
  - **lr = 5e-5** → F1 Hierarchical = 0.62150



Model	lr	Weight Balancing	F1 Hierarchical	Precision	Recall
BERT Model	$lr = 2e^{-5}$	NO	0.57082	0.70921	0.47763
BERT Model	$lr = 2e^{-5}$	YES	0.61390	0.58115	0.65057
BERT Model	$lr = 5e^{-5}$	YES	0.61996	0.57060	0.67867
BERT Model	$lr = 1e^{-4}$	YES	0.58580	0.55670	0.61811
BERT Model Hierarchical	$lr = 2e^{-5}$	YES	0.62071	0.54568	0.71967
BERT Model Hierarchical	$lr = 5e^{-5}$	YES	0.62150	0.58086	0.66826

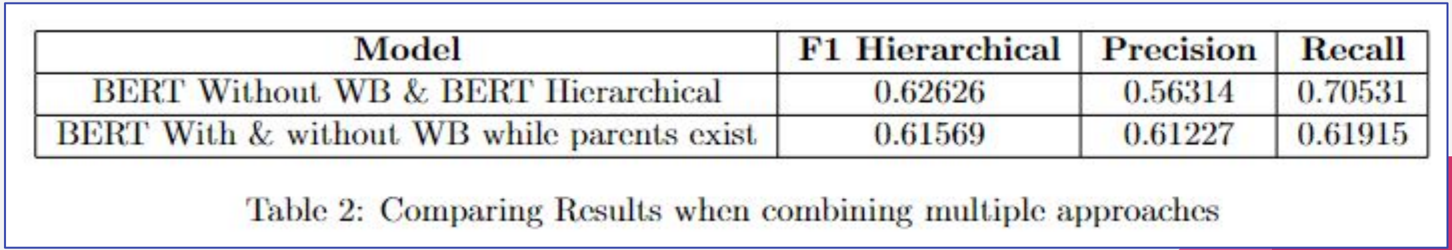
Table 1: BERT Results for different values of learning rates and different approaches (Hierarchical vs non-Hierarchical)

Combining BERT without weight balancing and Hierarchical BERT

- F1 Hierarchical = 0.62626
  ⇒ Better results when we compare it to the previously implemented hierarchical approach

Combining BERT with weight balancing and BERT Without Weight Balancing while parents exist

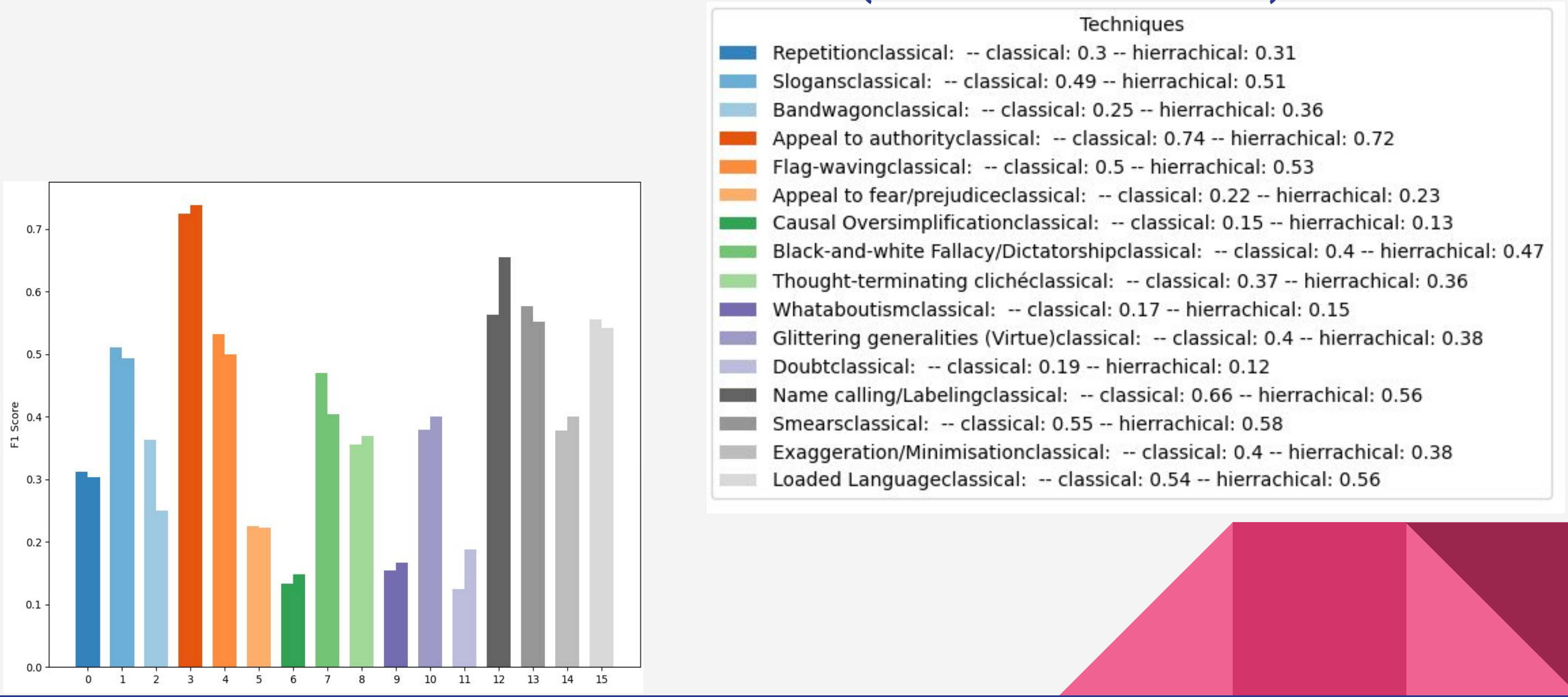
- F1 Hierarchical = 0.61569
  ⇒ We expected to obtain better results since the model that detects the ancestor nodes has a good performance (F1 Micro = 0.6557). However, the results show that this approach is less performant than the previous one.



Model	F1 Hierarchical	Precision	Recall
BERT Without WB & BERT Hierarchical	0.62626	0.56314	0.70531
BERT With & without WB while parents exist	0.61569	0.61227	0.61915

Table 2: Comparing Results when combining multiple approaches

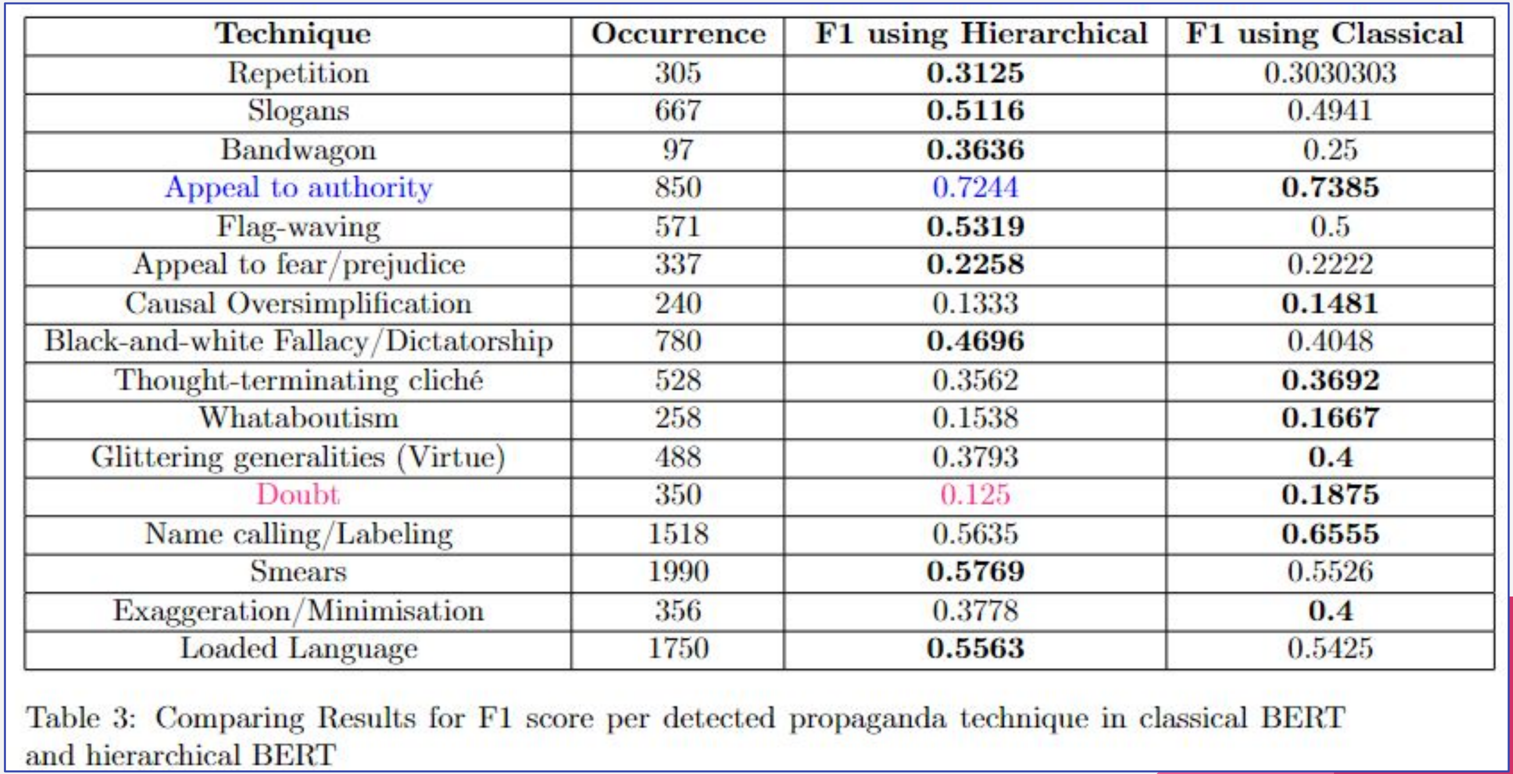
F1 Score Per Detected Propaganda Technique For Classical And Hierarchical BERT (Our Best Model)



Techniques

- 🟦 Repetitionclassical: -- classical: 0.3 -- hierrachical: 0.31
- 🟦 Slogansclassical: -- classical: 0.49 -- hierrachical: 0.51
- 🟦 Bandwagonclassical: -- classical: 0.25 -- hierrachical: 0.36
- 🟧 Appeal to authorityclassical: -- classical: 0.74 -- hierrachical: 0.72
- 🟧 Flag-wavingclassical: -- classical: 0.5 -- hierrachical: 0.53
- 🟧 Appeal to fear/prejudiceclassical: -- classical: 0.22 -- hierrachical: 0.23
- 🟩 Causal Oversimplificationclassical: -- classical: 0.15 -- hierrachical: 0.13
- 🟩 Black-and-white Fallacy/Dictatorshipclassical: -- classical: 0.4 -- hierrachical: 0.47
- 🟩 Thought-terminating clichéclassical: -- classical: 0.37 -- hierrachical: 0.36
- 🟪 Whataboutismclassical: -- classical: 0.17 -- hierrachical: 0.15
- 🟪 Glittering generalities (Virtue)classical: -- classical: 0.4 -- hierrachical: 0.38
- 🟪 Doubtclassical: -- classical: 0.19 -- hierrachical: 0.12
- ⬛ Name calling/Labelingclassical: -- classical: 0.66 -- hierrachical: 0.56
- ⬛ Smearsclassical: -- classical: 0.55 -- hierrachical: 0.58
- ⬜ Exaggeration/Minimisationclassical: -- classical: 0.4 -- hierrachical: 0.38
- ⬜ Loaded Languageclassical: -- classical: 0.54 -- hierrachical: 0.56

F1 Score Per Detected Propaganda Technique For Classical And Hierarchical BERT (Our Best Model)



Technique	**Occurrence**	**F1 using Hierarchical**	**F1 using Classical**
Repetition	305	**0.3125**	0.3030303
Slogans	667	**0.5116**	0.4941
Bandwagon	97	**0.3636**	0.25
Appeal to authority	850	0.7244	**0.7385**
Flag-waving	571	**0.5319**	0.5
Appeal to fear/prejudice	337	**0.2258**	0.2222
Causal Oversimplification	240	0.1333	**0.1481**
Black-and-white Fallacy/Dictatorship	780	**0.4696**	0.4048
Thought-terminating cliché	528	0.3562	**0.3692**
Whataboutism	258	0.1538	**0.1667**
Glittering generalities (Virtue)	488	0.3793	**0.4**
Doubt	350	0.125	**0.1875**
Name calling/Labeling	1518	0.5635	**0.6555**
Smears	1990	**0.5769**	0.5526
Exaggeration/Minimisation	356	0.3778	**0.4**
Loaded Language	1750	**0.5563**	0.5425

Table 3: Comparing Results for F1 score per detected propaganda technique in classical BERT and hierarchical BERT

Results Interpretation

- Weight balancing improves the results of BERT because it addresses class imbalance in the dataset ⇒ It gives more importance to under-represented classes and helps the model learn from these classes more effectively.

- Training BERT model on a hierarchy of 28 persuasion techniques, including parent categories, improves results because it allows the model to accurately recognize broader categories even when it can't identify specific techniques.

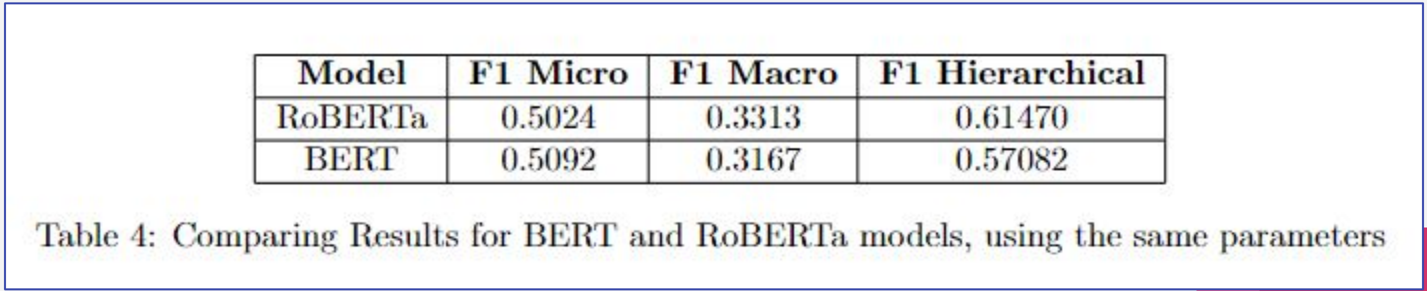
- By its nature and pre-training algorithm, we can notice that BERT model is performant in
  - Detecting the “Appeal to authority” technique even though it’s not the most frequent technique. ⇒ The best F1 score for both Classical and Hierarchical BERT.
  - Struggling in detecting the “Doubt” technique which has the lowest F1 score in Hierarchical BERT which is our best model.

COMPARATIVE ANALYSIS OF NLP MODELS

Comparative Analysis Of BERT & RoBERTa Models

ROBERTa: We used **RobertaForSequenceClassification**, a new version of the BERT model that removes the next-sentence prediction objective and training with much larger mini-batches and learning rates.

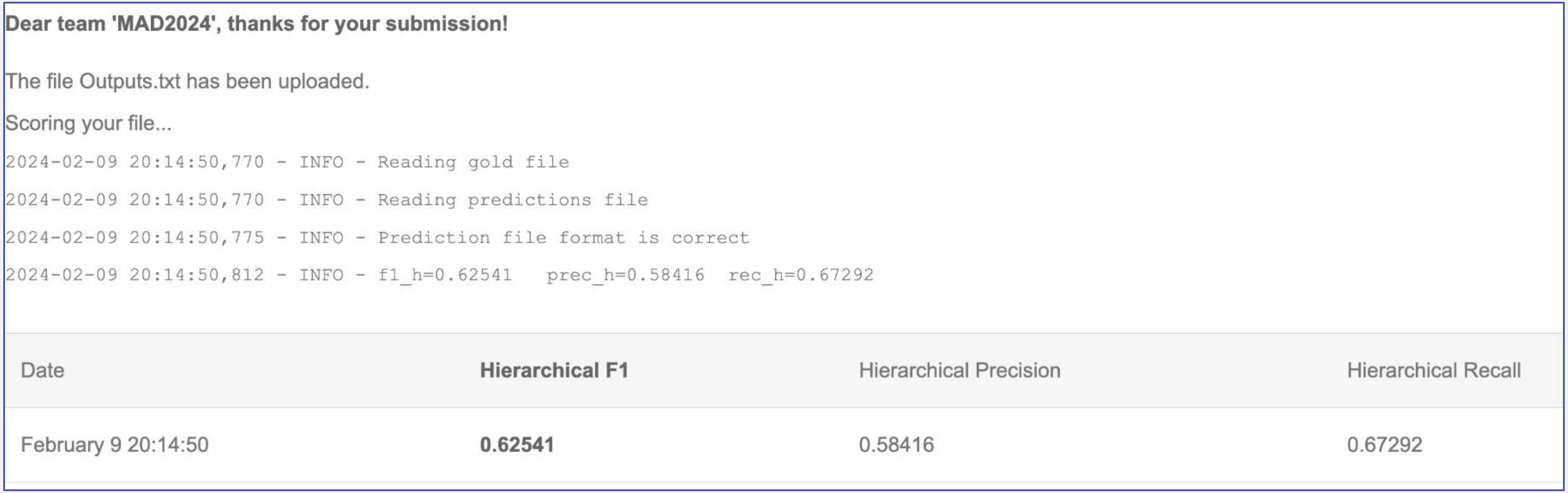
- **Results**



Model	F1 Micro	F1 Macro	F1 Hierarchical
RoBERTa	0.5024	0.3313	0.61470
BERT	0.5092	0.3167	0.57082

Table 4: Comparing Results for BERT and RoBERTa models, using the same parameters

FINAL SUBMISSION



Dear team 'MAD2024', thanks for your submission!

The file Outputs.txt has been uploaded.

Scoring your file...

```
2024-02-09 20:14:50,770 - INFO - Reading gold file
2024-02-09 20:14:50,770 - INFO - Reading predictions file
2024-02-09 20:14:50,775 - INFO - Prediction file format is correct
2024-02-09 20:14:50,812 - INFO - f1_h=0.62541   prec_h=0.58416  rec_h=0.67292
```

Date	**Hierarchical F1**	Hierarchical Precision	Hierarchical Recall
February 9 20:14:50	**0.62541**	0.58416	0.67292

CHALLENGES

- Limited prior knowledge of natural language processing

  ⇒ Implementation challenges.

- Model Performance: our initial approach using a 20-class model and binary cross-entropy with loss logits for loss calculation gave sub-optimal results.

CONCLUSION

- Understand and implement advanced NLP models (BERT and ROBERTA).

- Comparative analysis between different models and approaches + Opting for an optimization of BERT model.
- Adapting to the hierarchical nature of the data and refining model parameters

- Best model: the combination of Classical BERT without weight balancing & the Hierarchical BERT with weight balancing
    - The results: F1 Hierarchical = 0.62626 Precision = 0.5614 Recall = 0.70531

- Possibility of integrating visual features into multimodal meme data to complement textual analysis.

QUESTIONS ?