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

Meriem Abidi
EURECOM
Antibes, France
meriem.abidi@eurecom.fr

Meriem Dimassi
EURECOM
Antibes, France
meriem.dimassi@eurecom.fr

February 14, 2024

Abstract

In the age of rapid information dissemination, memes have emerged for both humor and disinformation. This study addresses the challenge set leveraged by SemEval 2024 Task 4: the multilingual detection of persuasion techniques in memes [2], focusing specifically on Sub task 1 - the identification of such techniques within meme text.

1 Introduction

Memes are increasingly used to disseminate information and misinformation online, drawing on a mixture of textual and visual elements to persuade and influence the public. The research ([4]) highlights that memes can effectively spread multi-modal propaganda, using a variety of rhetorical and psychological techniques. These techniques range from the exploitation of emotions to the use of logical fallacies and loaded language. Understanding and identifying these persuasive techniques in memes is crucial to mitigating misinformation, as it allows us to distinguish the intent and potential impact of content shared on social media platforms. SemEval 2024, the 18th International Workshop on Semantic Evaluation ([3]), aims to explore and evaluate the computational understanding of meaning in natural language. The event features a series of tasks designed to test various aspects of semantic analysis, including word sense identification, semantic parsing, co reference resolution and sentiment analysis, among others.
Our work focuses on the SemEval TASK 4 ([2]) "Multilingual Detection of Persuasion Techniques in Memes", and more specifically on subtask 1 which consists of a multi label classification challenge focused solely on the textual content of memes in order to identify which of the 20 organised persuasion techniques ([1]) are used in the text of a meme.
Based on the BERT Model ([7]), we explored a multi-label hierarchical classification framework to detect 20 distinct persuasion strategies. Our approach involved a comprehensive analysis of the data set provided, model experimentation and a nuanced adaptation of BERT to capture the hierarchical complexity of persuasion tactics. Although we encountered and overcame several implementation difficulties, our results highlight the potential and limitations of current NLP techniques in the nuanced field of meme analysis.

2 Related Work

2.1 A Survey on Computational Propaganda Detection

2.1.1 Focus of the study

The paper ([6]) focuses on Computational Propaganda which use automated or algorithm-driven techniques to manipulate and influence public opinion. This involves the use of digital tools, social media platform and algorithms to spread misinformation and disinformation.

Analysis Prospective: The propaganda detection is reviewed from a text analysis perspective.

Steps:

- Production of annotated datasets.
- Characterizing entire documents.
- Detecting the use of propaganda techniques at the span level.

2.1.2 Available Datasets

- **TSHP-17:** a balanced corpus with document-level annotation including 4 classes (Trusted, Satire, Hoax and Propaganda), 11 sources and 22,580 articles.
- **QProp:** an annotated dataset with information from MBFC. It associates different metadata to each article such as the bias level and geographical information, average sentiment, publication date, identifier, author, and official source name from GDELT.

⟶ **Limitations:** lack of information about the precise location of a propagandist snippet within a document. Based only on the binary propaganda.

- **PTC:** an annotated dataset that goes deeper into the types of propaganda, considering 18 propaganda techniques, rather than the binary propaganda vs non- propaganda setting.

⟶ **Limitation**: the volume of PTC is way lower than that of TSHP-17 and QProp

2.1.3 Used techniques

Binary classification using TSHP-17 and QPop corpora

- **Logistic regression:** a statistical model used for binary classification. It's particularly useful when you want to predict a binary outcome.
- **SVM:** particularly effective for solving binary classification problems, but they can also be extended to handle multi-class classification and regression tasks. They are known for their ability to find the optimal hyperplane that best separates data points into different classes.

Observation: It is possible to extend the model into predicting the source by using distant supervision in conjunction with rich representations.

Classification using PTC corporate

- Binary classification
- Multi-label multi-class classifications and span detection task

⟶ **BERT- based contextual representations:** gave the best-performing models for both tasks used.

2.1.4 Observations

Text is not the only way to convey propaganda, pictures should also be used because they may convey stronger messages. Explainability (explaining the reason why a text is considered as a propaganda or not) is a desirable feature. However, the use of deep learning in the recent propaganda detections, lacks of explainability.
The use of recent AI systems that understand and generate human language (neural language models) have made it difficult even for humans to detect synthetic text.

The vast majority of existing detectors are evaluated only on a single annotated dataset.
When dealing with user-generated data, ethical considerations are also important.
Identifying the intent behind a propaganda campaign requires analysis that goes beyond individual texts, involving classification of the social media users that contributed to injecting and spreading propaganda within a network. Thus, importation of detecting malicious coordination.

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

2.2.1 Focus of the study

In the paper [5], the task focuses on detecting propaganda techniques:

- Detecting the techniques in the text.
- Detecting the text spans where the techniques are used.
- Detecting techniques in the entire meme, i.e., both in the text and in the image.

In this study, an inventory of 22 techniques were used, among which, the first 20 are applicable to both text and images, while the last two, "Appeal to (Strong) Emotions" and "Transfer", are reserved for images.

2.2.2 Dataset

English memes were collected from personal Facebook accounts over several months in 2020 by following 26 public Facebook groups, which focus on politics, vaccines, COVID-19, and gender equality.
The final annotated dataset consists of 950 memes: 687 memes for training, 63 for development, and 200 for testing.

2.2.3 Annotation Process

The selected memes were then annotated using the 22 persuasion techniques.
PyBossa was used as an annotation platform.

Phase 1 - Filtering and Text Editing: Some collected memes needed to be removed since they didn't fit the definition “photograph style image with a short text on top of it”. Google Vision API was used to extract the text from the memes.

Phase 2 - Text Annotation: Given the list of propaganda techniques for text only annotation, the annotators were asked to identify which techniques appear in the text.

Phase 3 - Text Consolidation: This phase was essential for ensuring quality and served as an additional training opportunity for the entire team.

Phase 4 - Multimodal Annotation: The goal is to identify which of the 22 techniques appears in the meme.

Phase 5 - Multimodal Consolidation

2.2.4 Results per Subtask

Subtask 1 (Unimodal: Text) Transformers were quite popular, and among them, most commonly used was RoBERTa, followed by BERT. Some participants used learning models such as LSTM, CNN, and CRF in their final systems, while internally, Naıve Bayes and Random Forest were also tried.

Subtask 2 (Unimodal: Text) BERT dominated, while RoBERTa was much less popular.

Subtask 3 (Multimodal: Memes) Transformers were quite popular for text representation, with BERT dominating, but RoBERTa being quite popular as well. For the visual modality, the most common representations were variants of ResNet, but VGG16 and CNNs were also used.

3 Data Analysis

3.1 Objective

In the preliminary phase of our project, we conducted a comprehensive data analysis to understand the distribution of propagandistic and non-propagandistic content across our datasets. We examined both the training and validation datasets in Subtask1 ([2]).

To effectively convey our findings, we used two types of visual representations:

- **Pie charts** were employed to illustrate the binary proportions of propagandistic versus non-propagandistic techniques in each dataset. These charts served as a tool to observe the overall balance or imbalance between these two categories.

- **Bar charts** were employed to display the distribution of various persuasion techniques within the datasets. Through these bar charts, we were able to not only identify the most prevalent techniques but also observe the relative occurrence of each technique.

So, our main objectives were to:

- ▶ Determine the number of propagandistic and non-propagandistic memes samples for each of the subtask datasets.

- ▶ Determine the list of used persuasion techniques for each of the subtask datasets and their number of occurrences.

3.2 Subtask 1 - Quantifying the distribution of persuasion techniques

3.2.1 Subtask 1 - Train Dataset

Our pie chart shows that most of the memes in the Subtask 1 train dataset are propagandistic (81.9%), with a smaller portion being non-propagandistic (18.1%). This tells us that our dataset is mostly made up of memes that use persuasion techniques. This can be seen in **figure 1**

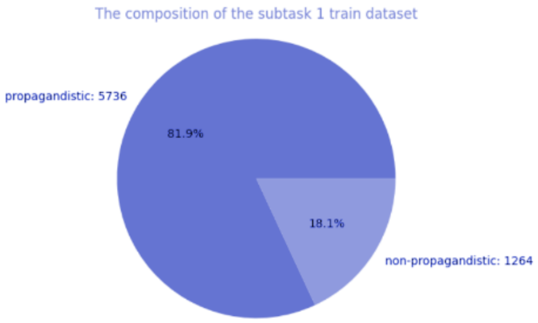

Figure 1: Subtask1 Train Dataset: Binary Classification

In our examination of the Subtask 1 train dataset, which is composed of 7000 samples, we observed a diverse distribution of the 20 persuasion techniques in meme texts. The technique of **'Smears'** is the most frequently occurring, with 1990 instances. **'Loaded Language'** follows with 1750 occurrences. 'Name Calling/Labeling' is also frequent, with 1518 instances. The 'Appeal to Authority' technique, with 850 instances, and 'Black-and-white Fallacy/Dictatorship', with 780 instances, present the top five techniques in the dataset.

On the other hand, the less frequent techniques such as 'Presenting Irrelevant Data (Red Herring)' and 'Obfuscation, Intentional Vagueness, Confusion' are observed to have 59 and 21 instances respectively, which suggests that these methods are less common in the context of the dataset. These distribution of persuasion techniques in the train dataset can be seen in **Figure 2**

3.2.2 Subtask 1 - Validation Dataset

In the validation dataset for Subtask 1, we observed a similar trend to what was found in the train dataset. The majority of samples were identified as propagandistic, constituting 82.4% of the dataset, while non-propagandistic content made up the remaining 17.6%. This binary classification is represented in **Figure 3**.

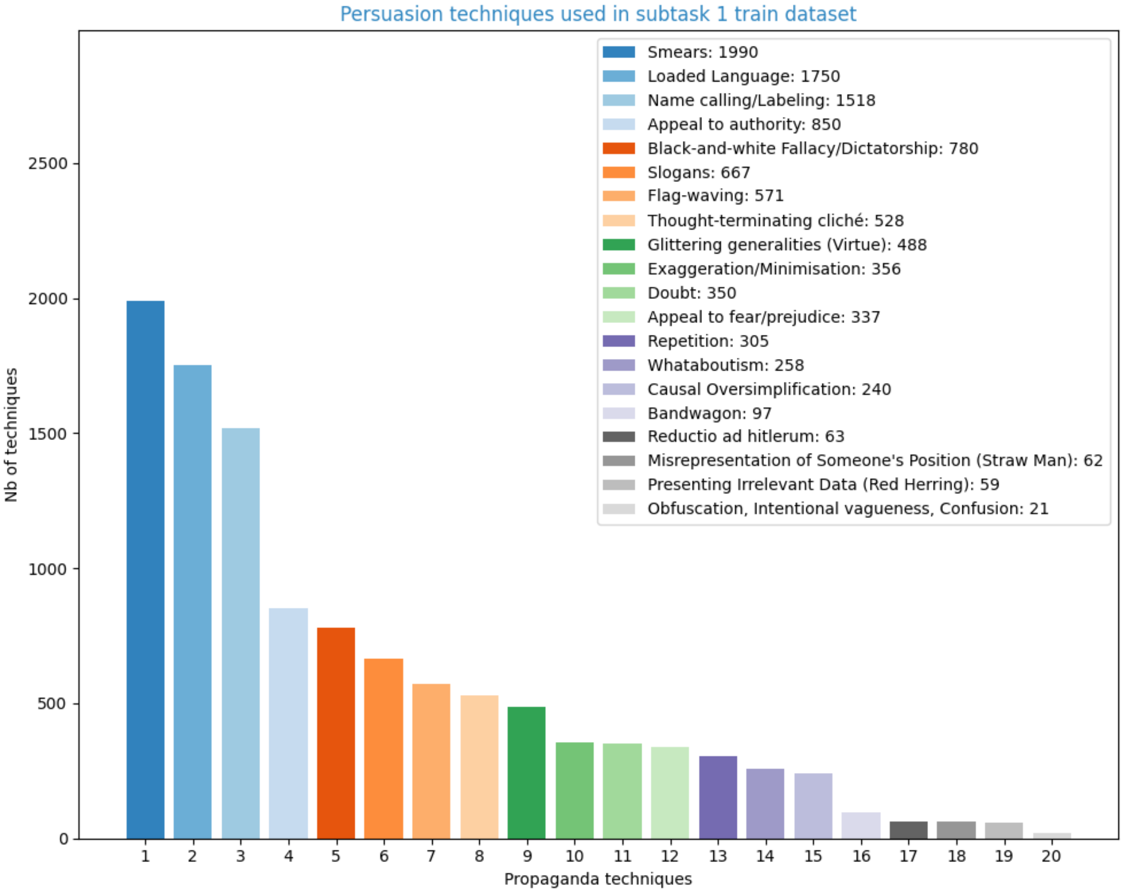

Figure 2: Subtask1 Train Dataset: Distribution of persuasion techniques

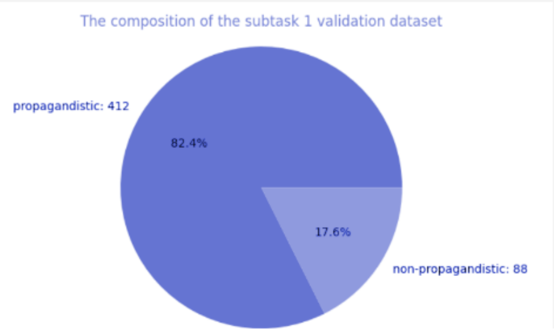

Figure 3: Subtask1 Validation Dataset: Binary Classification

The most prevalent techniques in the validation dataset are 'Smears', 'Loaded Language' and 'Name Calling/Labeling'. These represent the same top techniques observed in the train dataset. We can see the distribution in **Figure 4**

4 Proposed Approach

4.1 BERT Model

When implementing our model , we adapted BERT [7] to the specific task of detecting persuasision techniques in memes, which represented a complex challenge due to the multimodal nature of the data.

Pre-processing the textual content of memes: We extracted the text and assigned corresponding labels based on the presence of persuasive techniques. The labels were encoded using a *MultiLabelBinarizer*.

Model used: For the model, we used the *BertForSequenceClassification* from the *Hugging Face* transformation 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

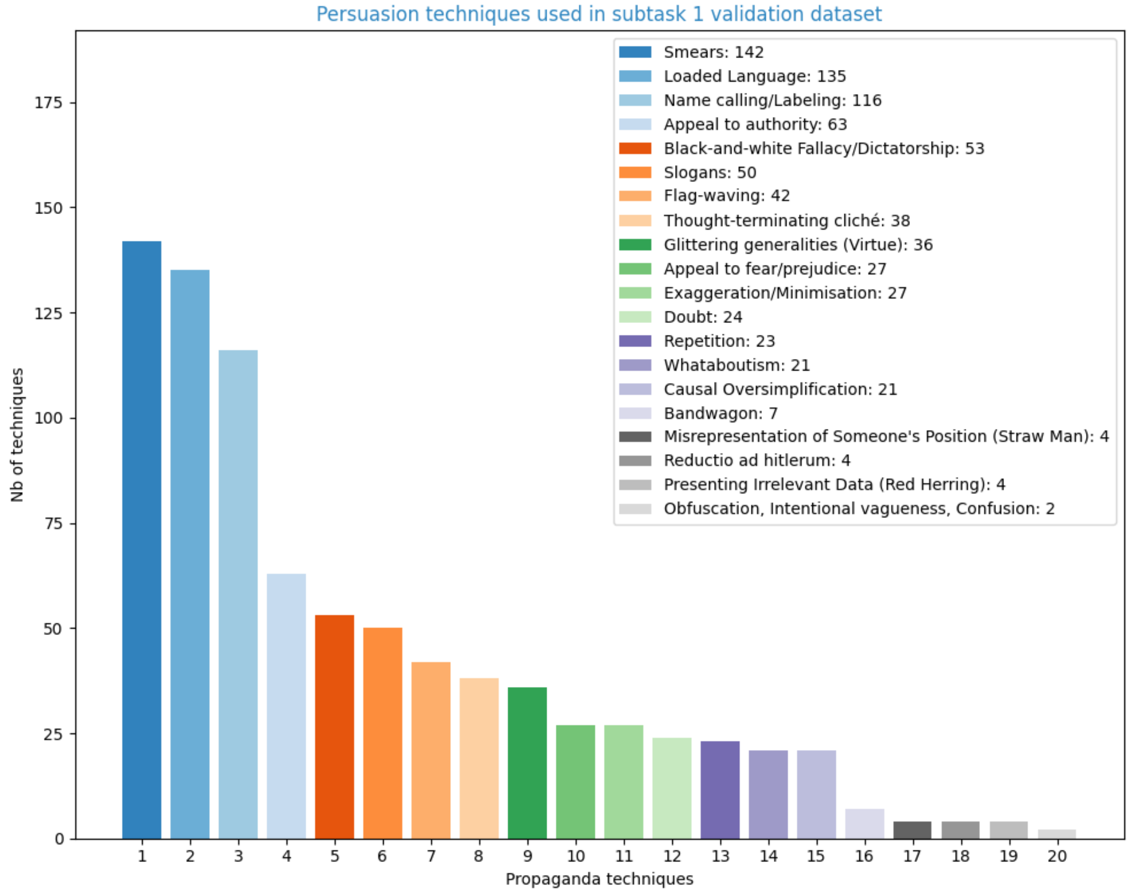

Figure 4: Subtask1 Validation Dataset: Distribution of persuasion techniques

"existence" of the technique.

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

Training phase: Involved introducing batches of tokenized text, attention masks and token identifiers over several epochs into the model.

Evaluation: During evaluation, predictions were generated for the validation dataset. These predictions, along with the true labels, were used to calculate various performance measures (F1 scores, precision and recall) to monitor the model's performance.

4.2 Optimization of Model Parameters

4.2.1 Number of training epochs

An insufficient number of training epochs can lead to under-fitting, as the model fails to capture the complexity of the data and therefore fails to adjust the model parameters. On the opposite, too many training epochs can lead to over-fitting, as the model learns the training data too well, including false predictions, leading to poor generalisation on unseen data.

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 accuracy.

4.2.2 Classes weights

Due to the variation in the frequency of different persuasion techniques in the dataset (**Section 3**), class imbalance was a problem that could decrease the performance of the model.

To address this, we assigned weights to the classes that were inversely proportional to their frequency.

This method gives more weight to less frequent classes when calculating the loss, allowing the model to pay more attention to these classes and improving its ability to detect them.

4.2.3 Learning rate

The learning rate determines the step size at each iteration while moving towards a minimum of a loss function. A rate that is too high can cause the model to converge too quickly to a suboptimal solution, while a rate that is too low can cause the convergence time to be too long or the model to get stuck in local minima.
For our BERT model, we tried different learning rates in order to optimize the training process.

4.3 Data Hierarchical Adaptation

4.3.1 Data Hierarchy

This task is a multi-label hierarchical classification problem. The hierarchy is essentially a directed acyclic graph that groups subsets of techniques sharing similar characteristics into a hierarchical structure as shown in the **figure** 5

4.3.2 Hierarchical Scores

In the prediction results, if only the ancestor node (parent node) of an existing technique is selected, only a partial reward is given.
Thus, correctly detecting the parent without having to know the exact technique can lead to an improvement in the "hierarchical F1" used in the final ranking compared to not having detected the technique.

4.3.3 Our Hierarchical Approach

Knowing the hierarchy of persuasion techniques, our model's training strategy was specifically designed to respect the parent-child relationships between techniques.
Instead of working on the 20 techniques, we chose to train the model on 28 techniques, including parents. In doing so, we aimed to improve the model's ability to first identify larger categories of persuasion methods before focusing on more specific techniques.
As a result, if a child technique was not recognized with certainty, the model was still able to classify the meme in the parent technique, resulting in a partial reward for the hierarchical F1 score.

4.4 Combining Multiple Approaches

4.4.1 Combine BERT Without Weight Balancing and Hierarchical Approach

To improve our model's detection capabilities, we tested a combination of different strategies. We combined the following two models:

1. **Classical BERT without weight balancing**, which allows the model to prioritize the most frequent techniques and improve their detection.

2. **Hierarchical approach with weight balancing**, which focuses on less common techniques while adding the ancestor nodes to the list.

Our model will compare the F1 score per propaganda technique at the end of the training phase for both combined models and then :

- If the difference between the F1 scores for a technique is greater than 0.02, we consider the best performing model as the predictor for that class.

- If the difference is less than 0.02, we consider the prediction of both models.

The dual strategy consisted of using the strengths of both approaches: the unbalanced BERT model was used to predict common techniques, while the weight-balanced hierarchical model was used to predict the nuances of rarer techniques.

4.4.2 Combine BERT with Weight Balancing and BERT Without Weight Balancing While Parents Exist

As part of this approach, we began by implementing another **"parents detection model"**, which is a classical BERT model but only detects ancestor nodes.
Since these nodes are much more frequent than the other persuasion techniques, the model

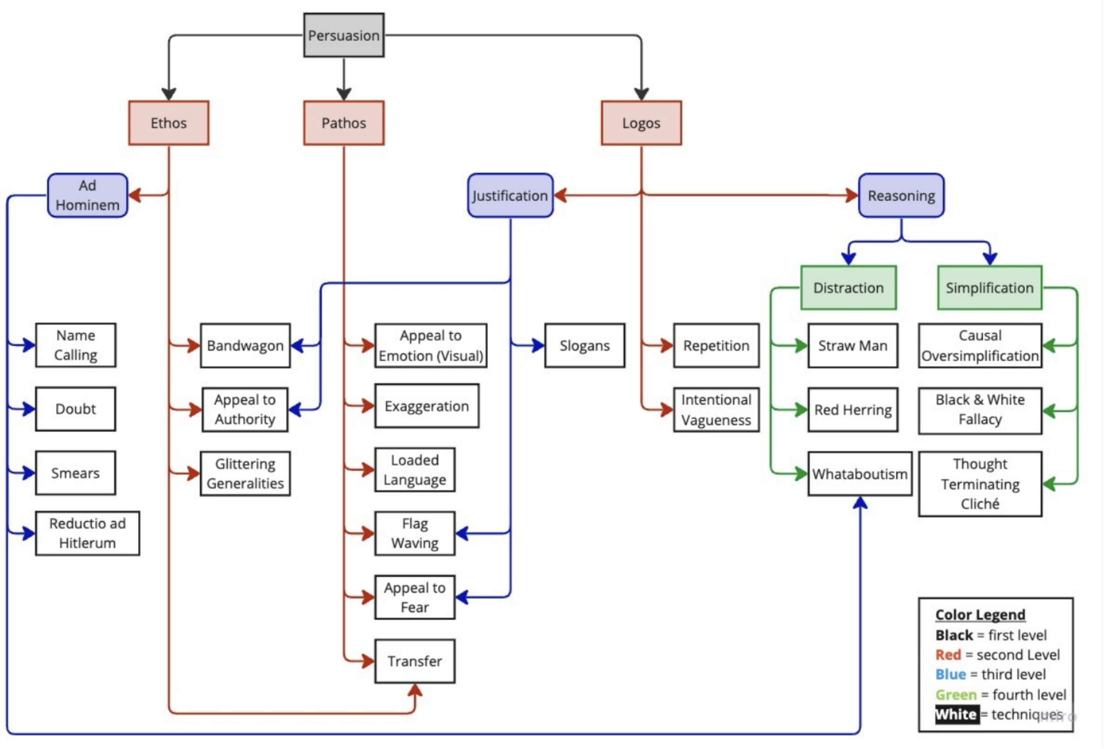

Figure 5: Hierarchy of the techniques for Subtask 2a (in Subtask 1 "Transfer" and "Appeal to Strong emotion" are not present).

should be able to detect these techniques more accurately.

This approach combines the results of:

1. Classical BERT model with weight balancing
2. Classical BERT model without weigh balancing

However, a technique is confirmed as predicted only if:

- it is identified by the most accurate model for that specific technique (similar to the previous approach)
- at the same time, one of its "parent" techniques is recognized by the parents detection model.

In doing so, we take advantage of the model's ability to detect ancestor nodes and this strategy should normally reduce incorrect predictions.

4.5 Comparative Analysis of NLP Models

Although various learning models such as LSTM, CNN, and CRF have been used for the detection of persuasive techniques, current state-of-the-art models for such NLP tasks mainly use BERT-based architectures .
Thus, in light of the analysis of SemEval-2021 Task 6 [5], which identified BERT and RoBERTa as the most commonly used transformers, we decided to continue our approach by a comparative analysis of the two models.

4.5.1 RoBERTa Model

For RoBERTa model, we used *RobertaForSequenceClassification* which is a new version of the BERT model that removes the next-sentence prediction objective and training with much larger mini-batches and learning rates. Researches ([8]) has proven that these changes are potentially able to improve performance in a variety of datasets. We maintained the same training and evaluation structure to do the compari-

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)

son with BERT model.

5 Results

5.1 Results for Parameter Optimization

5.1.1 Impact of Epochs Number

The evolution of the various metrics as a function of the number of epochs, from 1 to 10, is shown in **Figure 6** and **Figure 7**
We represented F1 Micro and F1 Macro for the BERT model with $lr = 2e^{-5}$ without weight balancing.

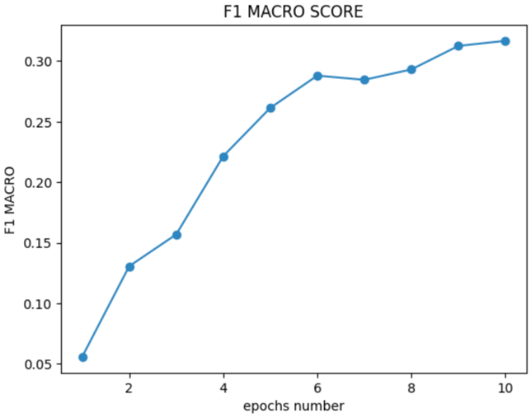

Figure 6: Evolution of F1 Micro as a function of number of epochs

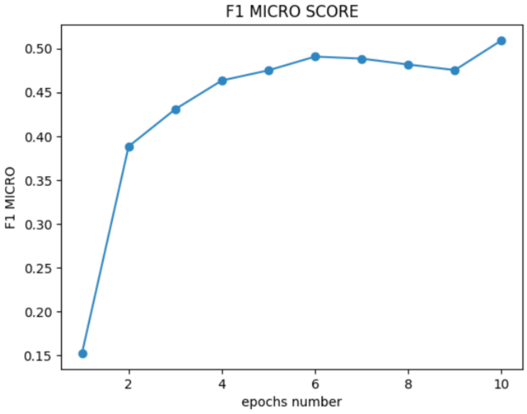

Figure 7: Evolution of F1 Macro as a function of number of epochs

5.1.2 Impact of Weight Balancing

In order to see the impact of weight balancing on the results, we have shown the metrics relative to the BERT model in the following two cases:

1. **No Weight Balancing:** this is the first approach that we used.
2. **Use of weight balancing:** by assigning weights to the different classes according to their occurrences.

As can be seen in **Table 1**, for the same learning rate $lr = 2e^{-5}$ the results for F1 Hierarchical have been improved, going from 0.57082 without weight balancing to 0.61390 when using weight balancing.

5.1.3 Impact of learning rate

Another approach we've used is to change the learning rate value. We obtained the metrics for the BERT model (using weight balancing) for 3 different values: $lr = 5e^{-5}$, $lr = 2e^{-5}$ and $lr = 1e^{-4}$.

As we can see in **Table 1**, the best F1 value obtained (0.61996) corresponds to $lr = 5e^{-5}$.

5.2 Data Hierarchical Adaptation

5.2.1 Our Hierarchical Approach

As discussed before **(Section 4.3.3)**, we implemented the BERT model with a hierarchical approach. We displayed the results for this approach for two different values of learning rate and we established a comparison. The main outcomes of this comparison are:

- For the same value of learning rate ($2e^{-5}$, the results of the hierarchical approach are better than the classical one. This can be seen through the increasing value of F1 Hierarchical that goes from 0.61390 to 0.62071.
- For the same hierarchical approach, the results obtained for $lr = 5e^{-5}$ are better than those obtained when using $lr = 2e^{-5}$ going from 0.62071 to 0.62150.

These results can be see in **Table 1**

5.2.2 Combining multiple approaches

Combine BERT without Weight Balancing and Hierarchical Approach

As mentioned before **(Section 4.4.1)**, we combined two different models:

1. Classical approach of BERT model without weight balancing
2. Hierarchical approach of BERT model with weight balancing

We displayed the results for this new approach. As we can see in **Table 2**, the first combination gives better results (F1 Hierarchical = 0.62626) when we compare it to the previously implemented hierarchical approach **(Section 4.3.3)**.

Combine BERT with Weight Balancing and BERT Without Weight Balancing while parents exist

As mentioned before **(Section 4.4.1)**, we combined two different models:

1. Classical approach of BERT model with weight balancing.
2. Classical approach of BERT model without weight balancing.
3. A technique is not taken into consideration unless one of its parents is detected.

We expected to obtain better results when using this approach, 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 performing than the previous one.

5.3 F1 score by propaganda technique

In **Figure 8** and **Table 3** we only represent the techniques that our models were able to detect. However for the following techniques, Our models gave an F1 score equal to 0:

- Presenting Irrelevant Data (Red Herring)
- Misrepresentation of Someone's Position (Straw Man)
- Obfuscation, Intentional vagueness, Confusion
- Reductio ad hitlerum

By its nature and its pre-training algorithm, we can see that the BERT model:

- Performs well in detecting certain persuasion techniques, such as "Appeals to authority", even though it is not the most frequent technique.

  It is better detected by both models, with F1 scores of 0.7244 for hierarchical BERT and 0.7385 for classical BERT.
- Struggles in detecting certain techniques despite their high frequency.

  For example, "Doubt" has an occurrence of 350, which is average, and still has the lowest F1 score in hierarchical BERT (0.125), as well as a low score in classical BERT (0.1875).

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

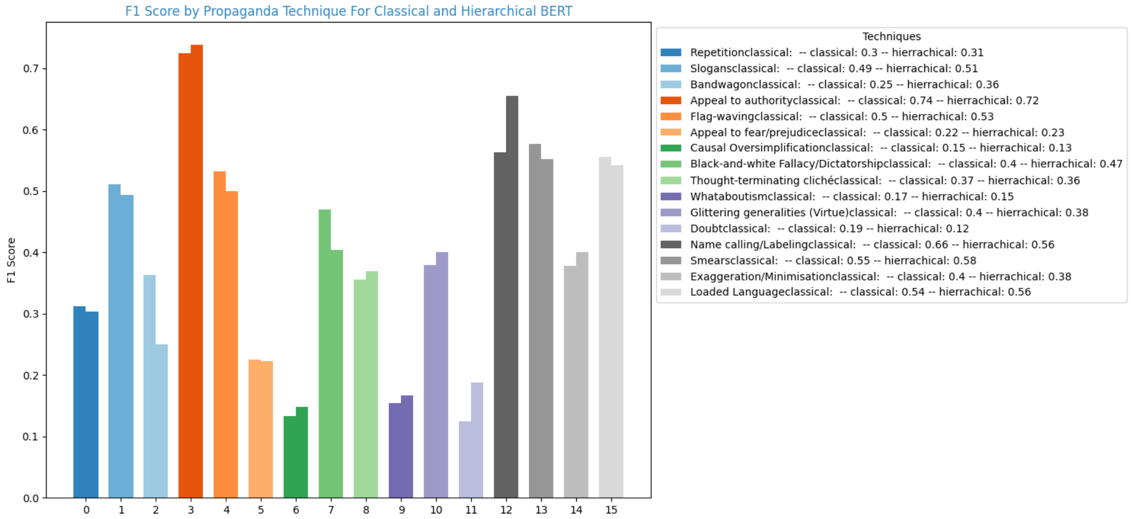

Figure 8: Graph representation of F1 score per detected propgande

Due to the difference in occurrence of the techniques :

- For techniques with a high number of occurrences, such as "Name call/labeling" and "Smears", the classical BERT, which does not use the weight-balancing technique, consistently displays higher F1 scores.

  This could mean this model is more robust in handling techniques that frequently appear in the data.

- The hierarchical model using weight balancing considerably improves the performance of the various techniques.

  For example, "Bandwagon" (occurrence = 97, considered low) has an F1 score of 0.3636 using hierarchical BERT and 0.25 using classical BERT.

⟶ By taking into consideration these two factors, hierarchical model which represents the most performing model, has the best F1 score for the technique "**Appeal to authority**" 0.7244 and struggles the most in detecting "**Doubt**" technique.

5.4 Results of the Comparative Analysis of NLP Models

We generated two models, BERT and RoBERTa, with the same parameters (Learning rate $= 2e^{-5}$ and without weight balancing), so that we could carry out a comparative analysis between the two models.

As can be seen in **Table** 4, through the values of F1 Micro, F1 Macro and F1 Hierarchical, the results of RoBERTa model are overall better than those of BERT.

RoBERTa outperformed BERT in detecting persuasion techniques in memes primarily due to its more extensive and diverse training data and optimizations in its training process such as the removal of the Next Sentence Prediction task.

6 Challenges and Solutions

6.1 Implementation Challenges

With limited prior knowledge of natural language processing, the implementation phase of the BERT model was slower than expected.

We spent a considerable amount of time understanding and effectively implementing the model. As a result, we did not have time to work on the other subtasks of SEMEVAL 2024 Task 4.

6.2 Model Performance

Our initial approach using a 20-class model and binary cross-entropy with loss logits for loss calculation gave sub-optimal results.
We spent a lot of time adjusting the logit threshold to achieve performance gains. When this did not yield the desired improvements, we pivoted our strategy to a 40-class model to better capture the nuances of our data.

7 Conclusion and Future Work

Throughout this project, we have attempted to make significant progress in the understanding and implementation of advanced NLP models, in particular BERT and RoBERTa.
A comparative analysis between different models and approaches was carried out, while opting for an optimization of these models. In most cases, this improved the ability to detect persuasive techniques in memes.
By adapting to the hierarchical nature of the data and refining model parameters, we have also tried to establish a solid basis for accurate classification.
Our best model, based on the value of F1 Hierarchical, is the combination of the Classical BERT without weight balancing **section 4.1** and the Hierarchical BERT with weight balancing (section 4.3.3).
The results of this model are the following:

- F1 Hierarchical = 0.62626
- Precision = 0.5614
- Recall = 0.70531

For the future, several possibilities are conceivable to advance this research. One possibility is to explore the integration of visual features into multi-modal meme data to complement textual analysis. Continuous refinement of model parameters and the exploration of new architectures may also further enhance performance.

References

[1] Semeval-2024: Persuasion techniques used in memes text. https://semeval.github.io/SemEval2024/.

[2] Semeval 2024 task 4 "multilingual detection of persuasion techniques in memes". https://propaganda.math.unipd.it/semeval2024task4/, 2023.

[3] Semeval-2024 the 18th international workshop on semantic evaluation. https://semeval.github.io/SemEval2024/, 2023.

[4] Shaden Shaar Firoj Alam Fabrizio Silvestri Hamed Firooz Preslav Nakov Dimitar Dimitrov, Bishr Bin Ali and Giovanni Da San Martino. Detecting propaganda techniques in memes. 2021.

[5] Shaden Shaar Firoj Alam Fabrizio Silvestri Hamed Firooz Preslav Nakov Dimitar Dimitrov, Bishr Bin Ali and Giovanni Da San Martino. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. 2021.

[6] Alberto Barron-Cedeno Seunghak Yu Roberto Di Pietro Giovanni Da San Martino1, Stefano Cresci2 and Preslav Nakov. A survey on computational propaganda detection. 2020.

[7] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[8] Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. Roberta: A robustly optimized bert pretraining approach. 2019.

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

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