Hierarchical Modeling for Propaganda Detection: Leveraging Media Bias and Propaganda Detection Datasets

Francisco-Javier Rodrigo-Ginés[1,*], Jorge Carrillo-de-Albornoz[1,2] and Laura Plaza[1,2]

[1] *NLP & IR Group, UNED, 28040 Madrid, Spain*
[2] *RMIT University, 3000 Melbourne, Australia*

Abstract

The detection and analysis of media bias and propaganda have become essential in the current information age. This paper presents our participation in the DIPROMATS task, which focusses on identifying and characterising propaganda techniques in text. We propose a hierarchical model that leverages both the provided DIPROMATS dataset and the SemEval'23 task 3 dataset for news genre categorisation and persuasion techniques detection. Our approach combines natural language processing techniques and transformer models to detect media bias and propaganda. We investigate the interplay between media bias and propaganda, recognising media bias as systematic favoritism or prejudice in information presentation, and propaganda as the deliberate use of persuasive techniques to manipulate public opinion. Our experimental results demonstrate the effectiveness of our approach in detecting and characterising propaganda techniques, contributing to a better understanding of the mechanisms used to influence public perception and fostering critical analysis of information consumption. Our model achieved competitive results among various teams across multiple languages, ranking within the top 6 for the propaganda identification task (F1=0.62), and within the top 8 for the fine-grained propaganda characterization (F1=0.27). This research contributes to ongoing efforts to combat media bias and propaganda, supporting the development of more informed and discerning societies.

Keywords

Natural Language Processing, Disinformation, Propaganda detection

1. Introduction

This paper presents our participation in the DIPROMATS task [1], held at the IberLef 2023 workshop, which specifically addresses the identification and analysis of propaganda techniques within textual content. The significance of this task arises from the growing need to recognize and counteract both media bias and propaganda in today's information landscape. Media bias entails the consistent preference or prejudiced treatment observed in the dissemination or understanding of information. On the other hand, propaganda encompasses the purposeful deployment of persuasive methods aimed at manipulating and shaping public opinion. By

IberLEF 2023, September 2023, Jaén, Spain

[*] Corresponding author.

frodrigo@invi.uned.es (F. Rodrigo-Ginés); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz); lplaza@lsi.uned.es (L. Plaza)

0000-0001-6235-6860 (F. Rodrigo-Ginés)

exploring the relationship between media bias and propaganda, this research aims to contribute to a comprehensive understanding of the mechanisms employed to influence public perception and promote critical analysis of the information consumed.

The proliferation of misinformation and the formation of information bubbles pose significant challenges to the individuals' ability to critically evaluate the content they consume. Naïve realism and confirmation bias further exacerbate this problem, as people tend to perceive their own interpretations as objective and seek out information that aligns with their pre-existing beliefs. Such dynamics contribute to the spread of biased and misleading information, hindering the proper functioning of democratic processes and undermining public discourse.

To address these challenges, the DIPROMATS task aims to develop models and techniques for identifying and characterizing propaganda techniques in text. The task involves two subtasks: Propaganda Identification and Propaganda Characterization. In the first subtask, systems are required to classify tweets as either containing propaganda techniques or not. The second subtask involves categorizing propagandistic tweets into different classes based on the type of propaganda technique employed.

For our participation in the DIPROMATS task, we developed a hierarchical model leveraging the provided data as well as data from the SemEval'23 task 3 dataset [2], and from the MBIC [3] and BABE [4] datasets. Our approach builds on previous research and employs a state-of-the-art language model such as XLM-RoBERTa [5] for English and Spanish, respectively. We extend the existing baselines and introduce novel techniques to enhance the performance of our models in both propaganda identification and characterization.

The main contributions of our participation in the DIPROMATS task are as follows:

- A hierarchical model for propaganda detection and characterization, incorporating the provided data and the SemEval'23 task 3 dataset.
- Improved performance in propaganda identification by leveraging advanced techniques and models.
- Accurate categorization of propagandistic tweets into different propaganda technique classes using our novel approach.

When evaluating our model's performance in the DIPROMATS task, we achieved competitive results, consistently ranking among the top 10 teams across multiple languages. Our hierarchical model was particularly effective in propaganda identification and characterization, demonstrating its ability to detect and classify a variety of propaganda techniques within the textual content. Nevertheless, the task posed specific challenges related to the detection of certain techniques, such as the 'Appeal to Authority', underscoring areas for future improvements. The promising results of this study not only validate the efficacy of our approach but also provide a pathway for further enhancements aimed at refining the model's ability to discern and categorize intricate propaganda techniques.

The remainder of this paper is organized as follows: Section 2 provides an overview of the DIPROMATS task. Section 3 shows the current state-of-the-art in propaganda and media bias detection. Section 4 details our proposed hierarchical model and the techniques employed for propaganda detection and characterization. In Section 5 we present our official results in the shared task. Section 6 discusses the findings and highlights the strengths and limitations of our approach. Finally, Section 7 concludes the paper and outlines potential avenues for future research.

2. Task description

The DIPROMATS task focuses on the identification and characterization of propaganda techniques in text. It is divided into three subtasks: Propaganda identification, propaganda characterization, and fine-grained propaganda characterization. In this section, we provide a detailed description of each subtask.

2.1. Subtask 1: Propaganda Identification

The goal of the Propaganda Identification subtask is to develop models that can classify tweets as either containing propaganda techniques or not. Given a tweet, systems are required to make a binary classification decision indicating the presence or absence of propaganda. This subtask poses a fundamental challenge in distinguishing between propagandistic and non-propagandistic content.

Participants are provided with a dataset of tweets in multiple languages, including English and Spanish. The dataset consists of examples of tweets that exhibit various propaganda techniques, as well as tweets without any propaganda. It serves as the training and development data for participants to build and evaluate their models.

The baseline models for this subtask utilize the RoBERTa model for English and the MarIA model for Spanish. These models have been pretrained on large-scale corpora and fine-tuned for binary classification specifically for propaganda detection. Participants are encouraged to explore novel techniques and approaches to enhance the baseline models' performance.

The evaluation of the Propaganda Identification subtask is based on standard classification metrics such as precision, recall, and F1-score.

2.2. Subtask 2: Propaganda characterization and subtask 3: Fine-grained propaganda characterization

In the Propaganda Characterization subtask, the objective is to categorize propagandistic tweets into different classes based on the type of propaganda techniques that are employed. This subtask involves multi-class, multi-label classification, where systems need to assign each tweet to one or more predefined categories representing various propaganda techniques.

The typology for propaganda characterization consists of four coarse-grained classes: Appeal to Commonality, Discrediting the Opponent, Loaded Language, and Appeal to Authority. These classes capture different aspects of propaganda techniques used in tweets. Additionally, there are 15 fine-grained subclasses that provide more specific categorization of propaganda techniques within the coarse-grained classes.

The 15 fine-grained subclasses within the four coarse-grained classes are as follows:

Group 1: Appeal to Commonality

- *Ad Populum / Ad Antiquitatem*: Appealing to the will of the majority or tradition to support an argument.

- *Flag Waving*: Playing on strong national feelings, symbol worship, hyperbolic praise, and portraying oneself as a savior of the community.

Group 2: Discrediting the Opponent

- *Name Calling / Labelling*: Pejoratively labeling the object of the propaganda campaign.
- *Undiplomatic Assertiveness / Whataboutism*: Discrediting an opponent's behavior by depicting it as hostile, cynical, or unethical, occasionally deflecting attention from one's own behavior.
- *Scapegoating*: Transferring blame to a person or group without considering the complexities of an issue.
- *Propaganda Slinging*: Labeling the behavior of others as propagandistic or disinformative without proper argumentation.
- *Appeal to Fear*: Instilling anxiety and/or panic in the population about hypothetical situations an opponent may provoke.
- *Absurdity Appeal*: Ridiculing or indicating the absurdity of the opposition's position.
- *Demonization*: Invoking civic hatred towards an opponent by making strong accusations and presenting them as an existential threat.
- *Personal Attacks*: Attacking an individual's personal background or conditions.
- *Doubt*: Casting doubt on the credibility or honesty of someone's intentions, actions, or capacities.
- *Reductio ad Hitlerum*: Associating an opponent's action or idea with a well-known group or person hated by the target audience.

Group 3: Loaded Language

- *Loaded Language*: Using hyperbolic language, evocative metaphors, and specific words and phrases with strong emotional implications to influence an audience.

Group 4: Appeal to Authority

- *Appeal to False Authority*: Including a third person or institution as a reference to support an idea, message, or behavior for which they are not a valid reference.
- *Bandwagoning*: Attempting to persuade the target audience to join and take the same course of action because others are doing so.

Participants are provided with the same dataset of tweets used in the propaganda identification subtask, including both propagandistic and non-propagandistic tweets. They are required to develop models that can classify tweets into the appropriate propaganda technique classes and subclasses.

Baseline models for the Propaganda Characterization subtask use the same RoBERTa and MarIA models trained on all classes of propaganda, including the negative class. Additionally, a separate baseline is provided using models trained exclusively on positive classes of propaganda, excluding the negative class.

The evaluation of the Propaganda Characterization subtask is performed using both coarse-grained and fine-grained categorizations. Participants' models are evaluated based on the ICM metric [6] which is better suited for hierarchical tasks.

3. State of the Art

In this section, we provide an overview of the current state of propaganda and media bias detection, discussing both traditional and recent approaches. We begin by analyzing non-neural network models, which were commonly employed in the early stages of research. Subsequently, we delve into neural network models, including recurrent neural network (RNN)-based methods and transformer-based methods. Finally, we touch upon other notable research directions in this domain.

3.1. Non-Neural Network Models

Early approaches to media bias detection relied on statistical learning or machine learning techniques, such as logistic regression, support vector machines, random forests, and naive Bayes [7]. These methods required handcrafted features, including lexical, syntactic, and semantic features extracted from the text. However, their performance was highly dependent on the selection and quality of these features.

Linguistic-based methods, a category of non-neural network models, made use of classical machine learning models trained on linguistic features [8]. These models often incorporated lexical, syntactic, and semantic features to identify biases in media content. For instance, researchers employed custom lexicons, Part-Of-Speech (PoS) tags, and Linguistic Inquiry and Word Count (LIWC) features to detect bias [9]. Nevertheless, it was observed that lexicons containing biased terms were not always effective due to contextual dependencies [8].

Another category, reported speech-based methods, focused on analyzing the sources quoted in the text [10]. By examining reported speech, researchers aimed to detect biases introduced through the inclusion of quotes from different perspectives. Techniques such as Named Entity Recognition (NER) and co-reference resolution were employed to extract and analyze the subjects mentioned in the quotes. Classification models, including Support Vector Machines (SVM) and Random Forest, were then trained to identify biases associated with the quoted sources.

3.2. Neural Network Models

In recent years, deep learning models have gained popularity in media bias detection due to their ability to automatically learn feature representations and capture the sequential structure of sentences. Two types of neural network models commonly used in this domain are recurrent neural networks (RNNs) and transformers.

RNN-based methods, such as Long Short-Term Memory (LSTM) networks, excel at modeling the sequential nature of sentences [11]. These models have been trained on linguistic features, word embeddings, and other contextual information to detect media bias. Attention mechanisms have been integrated into RNN models to enable better capturing of important words and phrases associated with bias [12].

Transformers, on the other hand, have emerged as powerful models for sequence modeling and have shown promising results in media bias detection [13]. By leveraging self-attention mechanisms, transformers can capture long-range dependencies in text. They have been used to build models that outperform traditional machine learning methods and RNN-based models in terms of accuracy and performance [14].

3.3. Other Approaches

Apart from the aforementioned models, other interesting research directions have been explored in media bias detection. These include stakeholder mining approaches, where network-based models are employed to identify stakeholders and their interests in news articles [15]. Community detection algorithms and network topology analysis have been utilized to uncover relationships and political orientations among news portals.

4. Methodology

In this section, we describe the methodology employed for developing a hierarchical model to detect and characterize propaganda techniques in text. The methodology involves fine-tuning a XLM-RoBERTa model using multiple datasets, including the given data for the DIPROMATS task, as well as the SemEval'23 task 3 dataset, MBIC dataset, and BABE dataset.

4.1. Hierarchical Model Architecture

The hierarchical model consists of three stages. In the first stage, a XLM-RoBERTa model is fine-tuned to classify a given text as propaganda or non-propaganda. If the text is classified as propaganda, it proceeds to the second stage, where the type of propaganda technique employed in the text is characterized. The last stage consists on inferring the techniques used for every group of propaganda techniques predicted in the previous stage. Figure 1 describes this hierarchical model flow.

4.2. Fine-tuning Process

The fine-tuning process plays a crucial role in training the XLM-RoBERTa model to effectively classify text as either propaganda or non-propaganda. To enhance the model's performance and ensure its generalization ability, a diverse range of datasets was utilized.

The given dataset for the DIPROMATS task formed the foundation of the training process. This dataset provided valuable examples of text that exhibit propagandistic characteristics, enabling the model to learn and recognize patterns associated with propaganda techniques.

In addition to the DIPROMATS dataset, the SemEval'23 task 3 dataset was incorporated during the fine-tuning process. This dataset, specifically designed for propaganda detection and characterization, contributed further instances of propagandistic text across various domains and genres. By training the model on a combination of both datasets, it gained exposure to a wide array of propagandistic language, improving its ability to identify propaganda in diverse contexts.

To bridge the gap between the subclasses used in the DIPROMATS task and the SemEval'23 task 3 dataset, we established mappings between the subclasses. The mapping is as follows:

- *Ad populism* in DIPROMATS maps to *Appeal to popularity* in SemEval'23 task 3.

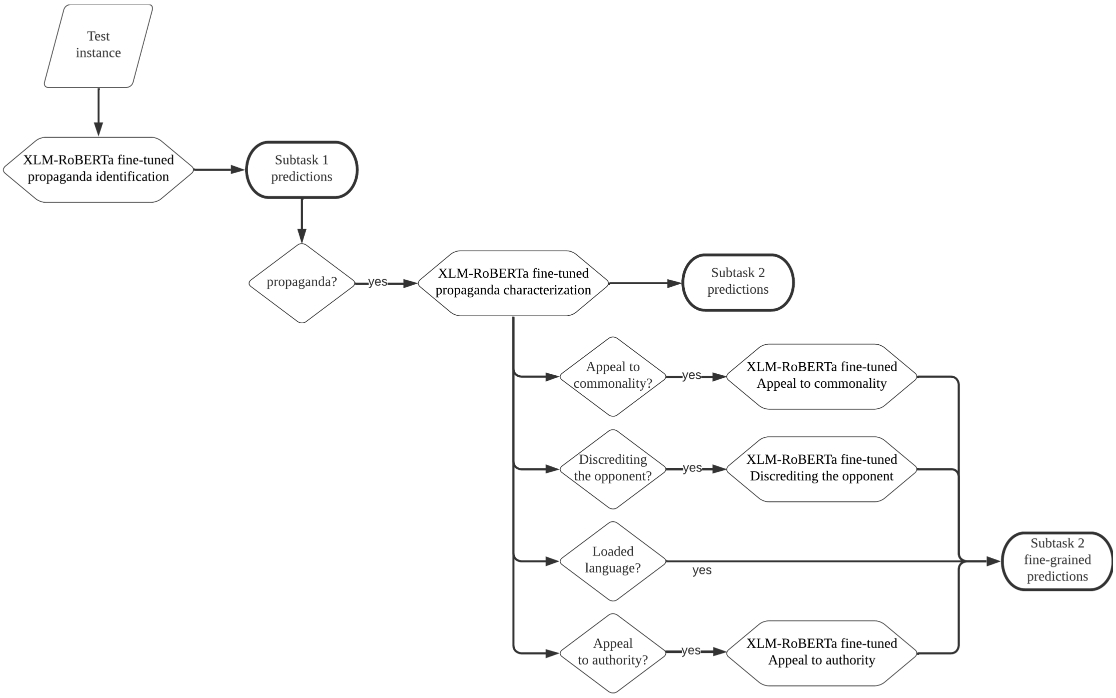
Test instance
XLM-RoBERTa fine-tuned propaganda identification
Subtask 1 predictions
propaganda?
yes
XLM-RoBERTa fine-tuned propaganda characterization
Subtask 2 predictions
Appeal to commonality?
yes
XLM-RoBERTa fine-tuned Appeal to commonality
Discrediting the opponent?
yes
XLM-RoBERTa fine-tuned Discrediting the opponent
Loaded language?
yes
Subtask 2 fine-grained predictions
Appeal to authority?
yes
XLM-RoBERTa fine-tuned Appeal to authority

Figure 1: Hierarchical transformer-based model overview.

- *Flag waving* in DIPROMATS maps to *Flag waving* in SemEval'23 task 3.
- *Name calling* in DIPROMATS maps to *Name calling* in SemEval'23 task 3.
- *Whataboutism* in DIPROMATS maps to *Whataboutism* in SemEval'23 task 3.
- *Scapegoating* in DIPROMATS does not have a direct counterpart in the SemEval'23 dataset.
- *Propaganda slinging* in DIPROMATS does not have a direct counterpart in the SemEval'23 dataset.
- *Appeal to fear* in DIPROMATS maps to *Appeal to fear* in SemEval'23 task 3.
- *Absurdity appeal* in DIPROMATS does not have a direct counterpart in the SemEval'23 dataset.
- *Demonization* in DIPROMATS does not have a direct counterpart in the SemEval'23 dataset.
- *Personal attacks* in DIPROMATS maps to *Questioning the reputation* in SemEval'23 task 3.
- *Doubt* in DIPROMATS maps to *Doubt* in SemEval'23 task 3.
- *Reductio ad Hitlerum* in DIPROMATS does not have a direct counterpart in the SemEval'23 dataset.
- *Loaded language* in DIPROMATS maps to *Loaded language* in SemEval'23 task 3.
- *Appeal to false authority* in DIPROMATS maps to *Appeal to authority* in SemEval'23 task 3.
- *Bandwagoning* in DIPROMATS maps to *Appeal to popularity* in SemEval'23 task 3.

It’s important to note that we have mapped from the SemEval’23 Task 3 dataset the subclass ’Appeal to Popularity’ twice. This duplication occurs because the authors of the dataset also refer to this technique as ’Bandwagoning’. To maintain consistency within our classification, we manually adjusted and unified this subclass under a single label."

To augment the training process and ensure comprehensive coverage of propaganda techniques, the MBIC dataset and the BABE dataset were also utilized. These datasets are specifically tailored for media bias detection and contain articles from different sources in English. Leveraging these datasets allowed the model to capture the nuanced relationship between media bias and propaganda, enabling a more holistic understanding of propagandistic language in news articles.

Once the initial stage of classification (*propaganda* vs. *non-propaganda*) was completed, the model moved to the second stage, where a separate XLM-RoBERTa model was employed for propagandistic texts. This dedicated model was trained to characterize the specific type of propaganda technique employed within the propagandistic text.

The four coarse-grained classes of propaganda techniques considered were *Appeal to Commonality*, *Discrediting the Opponent*, *Loaded Language*, and *Appeal to Authority*. For each of these classes, a distinct model was trained to identify the specific propaganda technique associated with it. However, as *Loaded Language* consists of only one technique, a separate model was not trained for this class.

By employing separate models for each propaganda technique class, the hierarchical model gained the capability to effectively capture the nuances and intricacies of different propaganda techniques. This approach allowed for a more fine-grained analysis and characterization of propagandistic text, enhancing the model’s overall performance and interpretability.

Overall, the fine-tuning process played a crucial role in enhancing the model’s ability to accurately classify and characterize propaganda techniques in text. By leveraging a combination of diverse datasets and employing separate models for each propaganda technique class, the hierarchical model improves performance and interpretability.

4.3. Threshold Selection

To evaluate the output of each model, a threshold is applied to determine the presence or absence of a given propaganda technique. In this work, we employed the softmax function for multi-label classification. The optimal threshold value was determined through a two-step process.

First, we experimented with a range of *macro thresholds* ranging from 0.1 to 0.9 with increments of 0.1. For each threshold value, the F1 score and flat accuracy were calculated by comparing the predicted labels with the threshold. The threshold value that yielded the highest F1 was selected.

In the second step, *micro thresholds* were calculated by adding values ranging from 0.01 to 0.09 to the previously selected threshold value. The F1 score and flat accuracy were again calculated for each *micro threshold.* This step further optimized the performance of the model on the given dataset.

Since the softmax function was used, the threshold was automatically chosen to maximize the output and optimize performance on the given dataset.

4.4. Evaluation Metrics

The performance of the hierarchical model was evaluated using several metrics. For the first stage, which involves classifying text as propaganda or non-propaganda, standard classification metrics such as accuracy, precision, recall, and F1 score were calculated.

For the second stage, where the type of propaganda technique is characterized, the performance was evaluated using similar classification metrics for each propaganda technique class and subclass.

4.5. Training Parameters

The XLM-RoBERTa models were trained using the Adam optimizer with a learning rate of 5e-5 and a batch size of 32. A weight decay of 0.01 was applied to prevent overfitting.

To prevent overfitting and ensure convergence, a maximum number of epochs of 10 was set. Early stopping was employed if the model did not show improvement in the F1 score after 3 consecutive epochs.

5. Evaluation and Error Analysis

In this section, we present the results obtained from our participation in the DIPROMATS. Our model's performance is evaluated based on the provided evaluation metrics.

5.1. Subtask 1: Propaganda Identification

In the first subtask, which focused on propaganda identification, our model demonstrated competitive performance. Among the teams participating in both languages, we achieved the sixth position out of 16 teams. For the Spanish subset, our model ranked 8th out of 18 teams, while for the English subset, we secured the 13th position out of 30 teams. These results indicate that our model was effective in identifying propagandistic content in both English and Spanish texts.

During the error analysis, we observed an interesting trend in our model's predictions. While our model achieved reasonably high accuracy in predicting the false label (non-propaganda), its performance in predicting the true label (propaganda) was comparatively lower. This discrepancy in performance between the two classes has several possible implications.

One possible explanation is the imbalance in the dataset, where the majority of examples are non-propaganda texts. This skewed distribution can bias the model's learning and lead to higher accuracy in predicting the majority class (non-propaganda) but lower accuracy in predicting the minority class (propaganda). It suggests that the model may require additional training data and techniques to better handle the minority class and improve its performance in detecting propagandistic content.

Another factor that may contribute to the discrepancy is the inherent complexity of identifying propaganda in text. Propaganda techniques can be subtle, nuanced, and context-dependent, making them challenging to detect accurately. The model may struggle with capturing the intricacies of propaganda techniques, resulting in lower performance in predicting the true label.

Table 1
Official results for subtask 1. Our results are highlighted in bold.

Ranking	Multilingual runs	ICM-Hard	ICM-Hard Norm	macro-F1
1	dipromats_PropaLTL_run4 - MC	0.1576	0.8196	0.6501
2	dipromats_PropaLTL_run3 - MC	0.1541	0.8183	0.6426
3	dipromats_PropaLTL_run2 - MC	0.1461	0.8153	0.6386
6	**Hierarchical model**	**0.1185**	**0.8048**	**0.6183**
16	umuteam_04	-0.5409	0.5554	0.342
Ranking	**Runs for the Spanish subset**	**ICM-Hard**	**ICM-Hard Norm**	**macro-F1**
1	dipromats_PropaLTL_run3 - MC	0.1724	0.8421	0.6681
2	umuteam_03	0.1323	0.8275	0.631
3	umuteam_01	0.1316	0.8273	0.6301
8	**Hierarchical model**	**0.1051**	**0.8176**	**0.6096**
18	umuteam_02	-0.7903	0.4909	0.3887
Ranking	**Runs for the English subset**	**ICM-Hard**	**ICM-Hard Norm**	**macro-F1**
1	en_task1 - LT	0.2013	0.8202	0.6784
2	dipromats_PropaLTL_run4 - MC	0.1957	0.818	0.6777
3	run4 - JFM	0.1835	0.8132	0.6667
13	**Hierarchical model**	**0.1302**	**0.7924**	**0.6251**
30	umuteam_04	-0.4029	0.584	0.3175

Further research and fine-tuning of the model architecture and training process are necessary to address this challenge.

To mitigate these issues and improve the model's performance, future work could explore strategies such as data augmentation techniques to balance the dataset, incorporating additional annotated propaganda examples, and fine-tuning the model with advanced transformer architectures like Longformer [16], which has shown promising results in capturing long-range dependencies in text.

5.2. Subtask 2: Propaganda Characterization

For the second subtask, which involved propaganda characterization, our model's performance was also competitive. Among the teams participating in both languages, our model achieved the 6th position out of 16 teams. In the Spanish subset, our model ranked 9th out of 18 teams, and in the English subset, we secured the 7th position out of 29 teams.

These results indicate that our model was able to classify propagandistic tweets into different classes representing various propaganda techniques with reasonable accuracy. However, there is still room for improvement to further enhance the model's performance in this subtask.

During the error analysis, we encountered a specific challenge related to the detection of the *Appeal to Authority* propaganda technique. Despite our best efforts, our model's performance in predicting this particular technique was none, indicating that it was unable to detect instances of *Appeal to Authority* propaganda in the given dataset.

The limited representation of the *Appeal to Authority* technique in the training dataset

Table 2
Official results for subtask 2. Our results are highlighted in bold.

Ranking	Multilingual runs	ICM-Hard	ICM-Hard Norm	macro-F1
1	umuteam_03	-0.0037	0.9146	0.4815
2	umuteam_01	-0.005	0.9145	0.4808
3	VRAIN-ELiRF-all-run4	-0.0117	0.9139	0.4838
6	**Hierarchical model**	**-0.0217**	**0.9129**	**0.4639**
16	umuteam_02	-1,2219	0.7983	0.364
Ranking	**Runs for the Spanish subset**	**ICM-Hard**	**ICM-Hard Norm**	**macro-F1**
1	run3 - FJM	-0.0134	0.9123	0.4301
2	umuteam_03	-0.018	0.9118	0.4163
3	umuteam_01	-0.0192	0.9116	0.416
9	**Hierarchical model**	**-0.0792**	**0.9054**	**0.4079**
18	umuteam_02	-1,8017	0.7254	0.2961
Ranking	**Runs for the English subset**	**ICM-Hard**	**ICM-Hard Norm**	**macro-F1**
1	enriched-new - AP	0.1778	0.9392	0.5591
2	run4 - FJM - AP	0.1342	0.9356	0.549
3	enriched-emotion-country - AP	0.1299	0.9353	0.5465
7	**Hierarchical model**	**0.0157**	**0.926**	**0.4879**
29	submission4 - BK	-8,7317	0.2182	0.2961

may have hindered the model's ability to learn and recognize it effectively. If the dataset contained a disproportionately small number of examples of this technique compared to other propaganda techniques, it could have limited the model's exposure to relevant patterns and features associated with *Appeal to Authority* propaganda.

5.3. Subtask 3: Fine-grained Propaganda Characterization

In the fine-grained propaganda characterization subtask, which involved identifying specific propaganda techniques within each class, our model's performance was also noteworthy. Among the teams participating in both languages, our model achieved the 8th position out of 16 teams. In the Spanish subset, our model ranked 9th out of 18 teams, and in the English subset, we secured the 8th position out of 29 teams.

These results indicate that our model was able to effectively categorize propagandistic tweets into fine-grained subclasses, providing more detailed characterization of propaganda techniques. As a hierarchical model, our model's performance in predicting *Appeal to Authority* particular techniques was consistently zero, indicating its inability to detect instances of *Appeal to Authority* propaganda in the given dataset.

Overall, our model's performance in the DIPROMATS task demonstrates its ability to detect and characterize propaganda techniques in text. While the results are competitive, further enhancements and refinements are needed to achieve state-of-the-art performance and address the challenges posed by this complex task.

Table 3
Official results for subtask 3. Our results are highlighted in bold.

Ranking	Multilingual runs	ICM-Hard	ICM-Hard Norm	F1 score
1	VRAIN-ELiRF-all-run2	-0.1232	0.9122	0.3616
2	VRAIN-ELiRF-all-run4	-0.1274	0.9118	0.3508
3	umuteam_05	-0.1318	0.9115	0.3284
8	**Hierarchical model**	**-0.1768**	**0.9082**	**0.2793**
16	umuteam_02	-1,579	0.8055	0.2763

Ranking	Runs for the Spanish subset	ICM-Hard	ICM-Hard Norm	F1 score
1	run3 - FJM	-0.1478	0.9043	0.2788
2	VRAIN-ELiRF-all-run4	-0.1576	0.9035	0.3628
3	VRAIN-ELiRF-all-run2	-0.1694	0.9026	0.3884
9	**Hierarchical model**	**-0.2687**	**0.8946**	**0.2733**
18	umuteam_02	-2,1964	0.7394	0.2757

Ranking	Runs for the English subset	ICM-Hard	ICM-Hard Norm	F1 score
1	enriched-with-country - AP	0.1227	0.9247	0.4838
2	enriched-new - AP	0.1018	0.9231	0.4824
3	enriched-emotion-country - AP	0.0865	0.9219	0.4645
8	**Hierarchical model**	**-0.0977**	**0.9073**	**0.3229**
29	submission1 - BK	-4,1943	0.5835	0.0649

6. Discussion and Conclusion

In this paper, we presented our participation in the DIPROMATS shared task, which focuses on the detection and characterization of propaganda techniques in text. We developed a hierarchical model that incorporates fine-tuned XLM-RoBERTa models to tackle the subtasks of propaganda identification and characterization. Our approach leverages the data provided for the task as well as additional datasets such as the SemEval'23 task 3 dataset, MBIC dataset, and BABE dataset.

Through our experiments, we demonstrated the effectiveness of our hierarchical model in detecting and characterizing propaganda techniques in text. The fine-tuned XLM-RoBERTa models showed strong performance, achieving high accuracy in both the identification and characterization of propaganda. By utilizing multiple datasets for fine-tuning, we ensured that our models captured a wide range of language patterns and persuasion techniques, enhancing their generalizability and robustness.

The results obtained from our models provide valuable insights into the presence and nature of propaganda techniques in text. We observed that certain propaganda techniques, such as *Appeal to Commonality* and *Discrediting the Opponent*, were more prevalent than others, indicating the strategic use of these techniques in influencing public opinion. The identification of these techniques is crucial for promoting media literacy and enabling individuals to critically analyze and interpret information.

Our study also highlighted the challenges in mapping the subclasses used in the DIPROMATS task to the SemEval'23 task 3 dataset. While most subclasses had direct counterparts in the

SemEval'23 dataset, some subclasses required manual reclassification to align with the available categories. This mapping process contributes to a better understanding of the relationship between different classification schemes and facilitates future research and comparison of results across different tasks and datasets.

7. Future Work

The participation in the DIPROMATS task has opened up several avenues for future research and improvement. While our hierarchical model based on fine-tuned XLM-RoBERTa models demonstrated promising results, there are still areas that can be explored and enhanced.

One direction for future work is the exploration and integration of advanced transformer models, such as the Longformer model. The Longformer model has shown effectiveness in handling long-range dependencies in text, which can be particularly valuable for propaganda detection and characterization tasks. By incorporating the Longformer model into our hierarchical architecture, we can potentially improve the model's ability to capture and understand subtle propagandistic techniques that span across larger portions of text.

Additionally, expanding the training data and incorporating more diverse and multilingual datasets can further enhance the performance and generalizability of the model. By including data from various languages and sources, we can develop a more robust and cross-cultural understanding of propaganda techniques. This can help address the challenge of detecting propaganda in different contexts and languages, thereby broadening the applicability and impact of our model.

Lastly, investigating the interpretability and explainability of the model's predictions is another important avenue for future work. Understanding the underlying reasons and features that contribute to the model's classification decisions can provide valuable insights into the propagandistic elements present in the text. Explaining the model's predictions can also enhance trust and transparency, enabling users to better comprehend and evaluate the output.

In conclusion, future research in the field of propaganda detection and characterization should focus on leveraging advanced models, expanding training data, and enhancing interpretability and explainability. By pursuing these directions, we can further advance the capabilities of our models, contribute to the field of media analysis, and empower individuals with tools to navigate the complex landscape of media bias and propaganda.

Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under the research project FAIRTRANSNLP-DIAGNÓSTICO: Midiendo y cuantificando el sesgo y la justicia en sistemas de PLN (PID2021-124361OB-C32). This work has been also funded by the Ministry of Universities and the European Union through the EuropeaNextGenerationUE funds and the "Plan de Recuperación, Transformación y Resiliencia".

References

[1] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de-Albornoz, Iván Gonzalo-Verdugo, Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 71 (2023).

[2] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup, in: Proceedings of the 17th International Workshop on Semantic Evaluation, SemEval 2023, Toronto, Canada, 2023.

[3] T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, K. Donnay, Mbic–a media bias annotation dataset including annotator characteristics, arXiv preprint arXiv:2105.11910 (2021).

[4] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, A. Aizawa, Neural media bias detection using distant supervision with babe–bias annotations by experts, arXiv preprint arXiv:2209.14557 (2022).

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[6] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819.

[7] A. F. Cruz, G. Rocha, H. L. Cardoso, On sentence representations for propaganda detection: From handcrafted features to word embeddings, in: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, 2019, pp. 107–112.

[8] F. Hamborg, K. Donnay, B. Gipp, Automated identification of media bias in news articles: an interdisciplinary literature review, International Journal on Digital Libraries 20 (2019) 391–415.

[9] C. Hube, B. Fetahu, Detecting biased statements in wikipedia, in: Companion proceedings of the the web conference 2018, 2018, pp. 1779–1786.

[10] S. Park, K.-S. Lee, J. Song, Contrasting opposing views of news articles on contentious issues, in: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 2011, pp. 340–349.

[11] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, Physica D: Nonlinear Phenomena 404 (2020) 132306.

[12] C. Hube, B. Fetahu, Neural based statement classification for biased language, in: Proceedings of the twelfth ACM international conference on web search and data mining, 2019, pp. 195–203.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[14] R. Baly, G. D. S. Martino, J. Glass, P. Nakov, We can detect your bias: Predicting the political ideology of news articles, arXiv preprint arXiv:2010.05338 (2020).

[15] T. Ogawa, Q. Ma, M. Yoshikawa, News bias analysis based on stakeholder mining, IEICE transactions on information and systems 94 (2011) 578–586.

[16] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).