A Transformer and Data Augmentation-based Approach for Propaganda Identification and Classification

Paula Sanchez-Checa[1], Lucía Gallego-Torres[1], Marco Antonio Sanchez-Escudero[1], Cosmin Petre[1] and Isabel Segura-Bedmar[1]

[1]*Universidad Carlos III de Madrid, Av. Universidad, 30, Leganés, 28911, Spain*

Abstract

Propaganda is a persuasion technique that aims to influence the way in which certain events are interpreted by the audience. Social media is a channel through which it is easy for such messages to spread. The aim of this work is the detection of propaganda in tweets in English and in Spanish, and a coarse-grained classification of those that are propagandistic. To make these classifications, we have used different BERT-like, GPT-like and XLNet-like models. For the propaganda identification task, the systems we have proposed have achieved a F1-Macro of 0.7906 for tweets in English, 0.7759 for tweets in Spanish, and 0.7813 for the multilingual approach. For task1b, our proposed solutions have achieved a F1-Macro of 0.3868 for tweets in English, 0.4628 for tweets in Spanish and 0.4486 for the multilingual approach.

Keywords

Propaganda identification, multilabel propaganda characterization, data augmentation, natural language processing

1. Introduction

Propaganda is not fundamentally dissimilar to advertising. It is simply a speech aimed at marketing an idea, whether it be political in nature or not. It is meant to persuade its target audience to believe something, whether it be true or false. Although propaganda can be found in many forms, the scope of this paper and of the DIPROMATS 2024 competition is to study propaganda in English and Spanish tweets authored by Chinese, Russian, US, and EU diplomats. As such, two main tasks have been proposed for DIPROMATS 2024: **propaganda identification and characterization (Task 1)** and **narrative detection (Task 2)**. Now, this paper tackles the first task, which is subdivided in three distinct subtasks:

- **Subtask 1a: Propaganda identification.** This subtask requires a binary classification system that can detect the use of propagandist methods in a given tweet.
- **Subtask 1b: Propaganda characterization, coarse-grained.** For this subtask a more complex system is necessary, a multiclass, multilabel classifier that can mark tweets as

IberLEF 2024, September 2024, Valladolid, Spain

paulasan@inf.uc3m.es (P. Sanchez-Checa); lucgalle@inf.uc3m.es (L. Gallego-Torres); marcoasa@inf.uc3m.es (M. A. Sanchez-Escudero); cpetre@pa.uc3m.es (C. Petre); isegura@inf.uc3m.es (I. Segura-Bedmar)
0009-0002-2616-687X (P. Sanchez-Checa); 0009-0004-8000-3064 (L. Gallego-Torres); 0009-0001-8163-5440 (M. A. Sanchez-Escudero); 0009-0008-5533-5103 (C. Petre); 0000-0002-7810-2360 (I. Segura-Bedmar)

belonging to these four categories: *Not Propagandistic*, *Appeal to Commonality*, *Discrediting the Opponent, Loaded Language.* Such a system would provide a deeper understanding of any detected propaganda.

- **Subtask 1c: Propaganda characterization, fine grained.** This final subtask requires a similar system as the previous one, having more specific categories to describe the propaganda found in tweets: *Flag Waving*, *Ad Populum / Ad antiquitatem*, *Name Calling/Labelling*, *Undiplomatic Assertiveness / Whataboutism*, *Appeal to Fear*, *Doubt*, and *Loaded Language.*

In order to accomplish this task, we pursued the use of LLMs, both dedicated Spanish and English language models as well as multilingual LLMs. We also employed data augmentation techniques to balance the provided training dataset. Specifically, we paraphrased the given tweets to generate more data for the underrepresented classes in the training dataset. Additionally, a second round of paraphrasing was applied to enlarge the balanced dataset. After each round of paraphrasing we made sure to remove duplicate tweets.

2. Related Work

DIPROMATS 2023 competition [1] focused on the automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers by using NLP techniques. The task was divided into several subtasks, including the detection of specific propaganda techniques and the classification of messages according to their propaganda content. It was concluded that the use of pre-trained models, such as BERT and its variants, combined with additional contextual features, can significantly improve the accuracy in detecting propaganda techniques [1]. Approaches that integrate pragmatic and contextual features were proven to be particularly effective. The works of the best-performing groups of DIPROMATS 2023 are exhibited below:

LiantTian [2]: This team used a system that employed a RoBERTa model that was fine-tuned to the task of detecting propaganda in tweets. In addition to the message content, additional features describing the overall communicative context were taken into account.

- In task 1, the model achieved an Information Contrast Model (ICM) of 0.1835 and F1 of 0.666. The multi-label confusion matrix made the model accurate for most of the techniques, although it had difficulties with techniques such as *Loaded Language* and *Appeal to authority.*
- In task 2, the model obtained an ICM of 0.1299 and F1 of 0.5465, and showed limitations in capturing pragmatic features. This suggests that extended contextual features are needed to improve performance.
- In task 3, the model was effective in detecting *Discrediting the Opponent* techniques, obtaining an ICM of 0.101 and F1 of 0.482. The techniques used included RoBERTa for sequence classification, fitting the pre-trained model in the classification task.

PropaLTL [3]: This team described the use of BERT with auxiliary contextual features for tweet propaganda detection. The auxiliary contextual features used were the historical context of tweets and user metadata. The model was fitted with nested data, incorporating additional features such as discursive context and metadata. An example of such features can be the user's posting history, which allows determining advertising patterns. This team participated in all tasks in English and Spanish, and obtained the following results:

- Their results in task 1 were for both English and Spanish. They obtained an ICM of 0.195 and F1 of 0.677 for English, and an ICM of 0.172 and F1 of 0.668 for Spanish.
- In task 2, they obtained an ICM of 0.179 and F1 of 0.659.
- In task 3, they obtained an ICM of 0.091 and F1 of 0.514.

Results for the tasks in English and Spanish showed significant improvements in accuracy by including auxiliary contextual features, especially in detecting complex propaganda techniques (such as historical context of tweets and user metadata).

UniLeon-UniBO [4]: Similar to the first team, a RoBERTa-based model for the detection of English and Spanish propaganda was presented. The approach included fine-tuning a RoBERTa model for tweets in both languages (Spanish and English). They obtained the following results in the 3 tasks:

- Task 1: High accuracy in detecting specific propaganda techniques in English. The results were an ICM of 0.134 and F1 of 0.549.
- Task 2: They only obtained results in English. Their ICM was of 0.069 and the F1 of 0.440.
- Task 3: The use of RoBERTa showed robustness in classifying multiple propaganda techniques in both languages, but improvement is recommended for Spanish because the obtained results were an ICM of -0.147 and F1 of 0.440.

LXMJ [5]: This group performed a cascade-based approach to detection language modeling. It consisted of a cascade of language models, including pre-trained models such as BERT and RoBERTa, fine-tuned specifically for the propaganda detection task. The cascade allowed faster and more accurate classification, taking advantage of the hierarchy of models such as BERT and RoBERTa. It proved to be efficient for propaganda detection compared to individual models, especially in tasks with high variability in propaganda techniques.

UMUTeam [6]: The UMUTeam combined linguistic features with contextual sentence embeddings to detect propaganda in English and Spanish. They used advanced LLMs, such as BETO and XLM, and fine-tuned them with task-specific data for the competition. The UMUTeam reported the following results about the tasks:

- Task 1: Good performance in both languages with an ICM of 0.132 and F1 of 0.631 when combining linguistic features with contextual embeddings.
- Task 2: The approach showed robustness in classifying more subtle advertising techniques. They only reported results in Spanish: ICM of -0.018 and F1 of 0.4164.
- Task 3: The combination of features improved propaganda detection in multilingual contexts. The results were only reported for the Spanish instances and were an ICM of -0.181 and F1 of 0.3414.

ELiRF-VRAIN [7]: They explored multilingual data augmentation to improve propaganda detection. They used multi-language data translation and synthesis techniques to improve model performance in both languages, which showed to improve model accuracy and robustness. This team participated in all tasks and obtained negative ICMs: ICM of -0.1576 and F1 of 0.362 for task1, ICM of -0.036 and F1 of 0.457 for task2 and ICM of -0.178 and F1 of 0.394 for the last task.

3. Methods

During the development phase, our methodology was based on three main steps: a pre-processing step, a model search step and a model validation step. The first step was pre-processing, where we studied the training dataset, cleaned the data, and balanced and augmented the dataset. The second step was the search and first training of the models. In this phase, we were limited by the computing resources available to us. In the validation step, the models were evaluated with different parameters to obtain the model configuration that gave the best results for our problem.

During the prediction and validation phases, we used a stratified version of the training dataset to evaluate the models. We divided the training dataset into three parts: a training split, an evaluation split and a test split. To train the models for task1a, we used 70% of the original training dataset for the training split, 15% for the evaluation split and 15% for the test split. For task1b, the training split consisted of 75% of the original training dataset, 12% for the evaluation split and 13% for the test split.

We used the training split to train the models. Then, the evaluation splits were used to evaluate the performance of each of the models and adjust the hyperparameters to obtain the best results. Finally, the test splits were used to evaluate the models one last time and choose the ones that showed the best performance for our rounds.

Once the selected models were trained, we designed a new methodology to obtain the predictions. First, we used the model selected for task1a to make the predictions. Those predictions that were labelled as propagandistic were also evaluated by a model trained for task1b to perform a coarse-grained classification. This model classified the propagandistic texts into one or more of the following three categories: *appeal to the commonality*, *Discrediting the Opponent* and/or *Loaded Language*.

3.1. Pre-processing

In the pre-processing phase, our first step was to clean the dataset. Here, we extracted the features we found relevant and removed the rest of the dataset to speed up processing and training. In our case, we used the *language*, *text*, *label_task1* and *label_task2* columns of the dataset.

The second step was to tokenize the text and labels of the dataset using model-specific tokenizers to prepare the data for the model.

The training dataset comprised approximately 14,528 texts, of which 8,408 were in English and 6,120 in Spanish. An analysis of the texts revealed that the average length of the texts was 36.8 tokens.

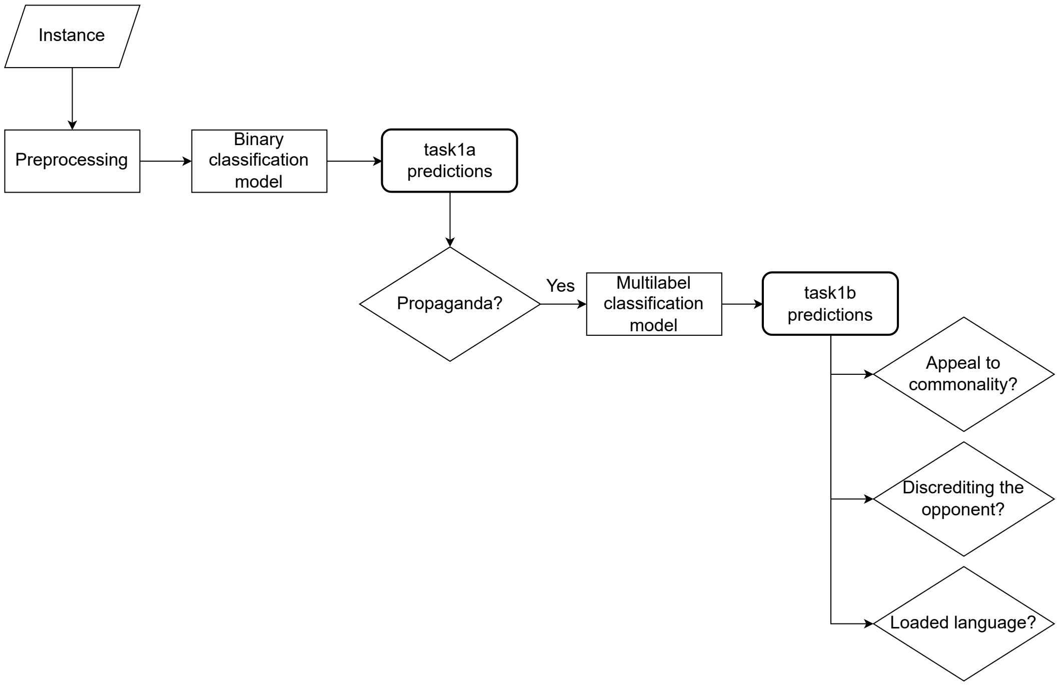

Figure 1: Methodology Diagram

Additionally, we set the maximum text length to 60 tokens. In the training dataset, the longest text had 66 tokens. In this way, most of the texts were complete and we optimized the model training phase.

A study on the distribution of classes in the training dataset showed that the false class was significantly more frequent than the true class, accounting for 76% and 24% of the instances in the dataset, respectively. In terms of the distribution of coarse-grained labels, the study showed that the label *Discrediting the Opponent* was significantly more represented in the dataset than the other two labels, accounting for 50% of the propagandistic instances. The label *Appeal to Commonality* accounted for 22% of the instances and the label *Loaded Language* for 28% of the cases.

One of the problems, we identified that could affect the predictions was the large imbalance in the dataset for the labels of task1a. This imbalance caused the models to have a high number of false negatives. To solve this problem, we balanced the dataset so that it contained the same number of instances with positive and negative labels. We used paraphrasing as a data augmentation technique to have a balanced dataset. We used a T5 model for paraphrasing English sentences that was pre-trained on the Google PAWS dataset [8]. We applied this paraphrasing model directly to the English dataset.

However, some additional steps for the Spanish paraphrasing were required. To use the same paraphrasing model as for the English dataset, we translated all the Spanish sentences into English. Then, the paraphrasing model was applied on them. Finally, we translated the generated sentences to Spanish. For the translation from Spanish to English, we used the model opus-mt-es-en, and for the translation from English to Spanish we used the model opus-mt-en-es. These two transformer models were developed in the Language Technology Research Group at the University of Helsinki [9] and were trained with the OPUS (Open Parallel Corpus) dataset.

Then, we checked that the new sentences generated after paraphrasing differed from the original sentences, and discarded all sentences that had not changed.

Moreover, we wanted to enlarge the dataset to check if we got better results when training the models with a larger amount of data. To do this, we used the data augmentation techniques we had used previously to balance the dataset over the balanced dataset. As we did when balancing the dataset, we discarded all paraphrased texts that did not change during the paraphrasing.

For task1b, due to the original unbalancing of task1a, by removing the instances labelled as non-propogandistic from the dataset, we had too little data to train the models. For this reason, we also carried out data augmentation techniques for this task. We used the same paraphrasing model as for task1a, and employed the double translation technique in order to be able to use the paraphrasing model in the Spanish texts.

3.2. Models

We have divided our search of LLMs into models trained for Spanish, English, and for both Spanish and English (referred to as multilingual from now on). There exist a wide variety of powerful LLMs for these languages, such as LLaMA-2 [10] and GPT2-large [11]. However, due to computation restrictions, we narrowed our search to those which did not require over 12.7 GB of RAM, 15 GB of GPU RAM or 201.2 GB of disk memory.

We trained the following LLMs with the training splits for task1a and task1b. We have chosen these models because, after a review of the literature and overviews of previous competitions, they were promising models for carrying out the text classification tasks. In addition, they all met our memory and GPU limitations.

- **BERT**: a transformer model developed by Google and first proposed in [12]. Traditionally, language algorithms read text sequentially, but BERT introduced the novelty of getting context from the surrounding text. We have used an uncased BERT model trained with data in English.
- **BETO**: a BERT-based model pre-trained on a Spanish corpus [13]. It is similar to the BERT-base model, differing in the corpus they were trained with.
- **RoBERTa**: a BERT-inspired model first proposed in [14]. RoBERTa offers an improved architecture and dynamic masking by using a larger training corpus, increasing the vocabulary and using larger batches of data. For the instances in English, we have used a RoBERTa-large model [15] pre-trained with raw English data. For the Spanish instances, we have used a RoBERTa-large model trained with data from the National Library of Spain [16].
- **RoBERTuito**: a BERT-based model pre-trained with a corpus consisting of tweets in Spanish [17, 18, 19].
- **RoBERTa-emoji**: a RoBERTa-base model pre-trained with Tweets and fine-tuned for emoji prediction [20].

- **GPT-2**: second GPT-based transformer model developed by OpenAI. This model has been pre-trained over a large dataset of data in English, consisting of webpages [11]. For the Spanish instances of the dataset, we used a GPT-2 model that was pre-trained with texts from the National Library of Spain [16].
- **XLNet**: is based on the Transformer-XL model and aims to learn bidirectional context for the predictions. In order to do so, XLNet maximizes the expected probability over all permutations of the factorization order of the input sequence by using an autoregressive approach[21].
- **XLM-Twitter**: a XLM-roBERTa-base model pre-trained on tweets in 8 languages, including English and Spanish, and fine-tuned for sentiment analysis [22].

The models were trained using a learning rate of 5e-5 and a weight decay of 0.01. Also, we used 500 warmup steps for training each model. Each model was trained for 4 epochs with a batch size of 32.

3.3. Threshold Selection

To evaluate the output of the models, we had to set thresholds for each of the tasks. The output of the models is the probability that an instance can be assigned with a label. Therefore, we had to establish from which probability we considered that an instance could be labelled.

For the binary classification of task1a, we used the softmax function to determine which label the model assigned to each instance, selecting the label with the highest probability.

task1b consisted of a multi-classification, in which instances could be labelled with one or more labels. In this case, we used a softmax function to determine the probability of each of the labels, and we manually set the threshold by which an instance is considered to be labelled with a label. This threshold was established by testing with thresholds from 0.1 to 0.9. The variations of the threshold were made in steps of 0.1. In each of these tests, we captured the F1-Macro obtained on the validation split, and selected the threshold with which we obtained the best F1-Macro. It is possible that sometimes none of the label probabilities exceeded the threshold, so in the case where no label exceeded the threshold, we used the softmax function to assign the label with the highest probability to the instance. The thresholds used for task1b can be seen in the Table 1.

4. Evaluation

4.1. Criteria for Submission Selection

For each task, we were permitted to submit up to five different runs. Each run could contain the predictions for English, Spanish, or both languages. We submitted all five runs for both Spanish and English texts, using different model combinations in each run. We used the following model combination for each run:

- **Run 1:**

Table 1
Task1b threshold configuration

model	dataset	task1b threshold
English		
RoBERTa-large	normal	0.2
	balanced	0.1
GPT-2-base	normal	0.1
	balanced	0.1
BERT-base-uncased	normal	0.3
	balanced	0.1
XLNet-base-cased	normal	0.1
	balanced	0.2
Spanish		
RoBERTa-large	normal	0.1
	balanced	0.1
GPT-2-base	normal	0.1
	balanced	0.1
BETO-base-uncased	normal	0.1
	balanced	0.1
RoBERTuito-sentiment-analysis	normal	0.1
	balanced	0.1
Multilingual		
XLM-Twitter	normal	0.3
	balanced	0.3
RoBERTa-emoji	normal	0.3

  - **task1a** (Language-specific models): RoBERTuito model trained with the original dataset for Spanish predictions and RoBERTa model trained with the augmented dataset for English predictions.
  - **task1b** (Language-specific models): RoBERTa model trained with the original dataset for Spanish predictions and XLNet model trained with the augmented dataset for English predictions.
- **Run 2:**
  - **task1a** (Language-specific models): RoBERTuito trained with the augmented dataset for Spanish predictions and XLNet model trained with the augmented dataset for English predictions.
  - **task1b** (Multilingual model): RoBERTa-emoji trained with the original dataset for both Spanish and English predictions.
- **Run 3:**
  - **task1a** (Language-specific models): RoBERTa model trained with the augmented dataset for Spanish predictions and BERT model trained with the augmented dataset for English predictions.

Table 2
Models employed in each run

	Model(s) for task1a		Model(s) for task1b	
Run	**Spanish**	**English**	**Spanish**	**English**
1	RoBERTuito Original dataset	RoBERTa Augmented dataset	RoBERTa Original dataset	XLNet Augmented dataset
2	RoBERTuito Augmented dataset	XLNet Augmented dataset	RoBERTa-emoji Original dataset	
3	RoBERTa Augmented dataset	BERT Augmented dataset	BETO Original dataset	RoBERTa Augmented dataset
4	RoBERTuito Original dataset	RoBERTa Augmented dataset	RoBERTuito Augmented dataset	BERT Augmented dataset
5	XLM-Twitter Original dataset		XLM-Twitter Augmented dataset	

  - **task1b** (Language-specific models): BETO model trained with the original dataset for Spanish predictions and RoBERTa model trained with the augmented dataset for English predictions.

- **Run 4:**
  - **task1a** (Language-specific models): RoBERTuito model trained with the original dataset for Spanish predictions (same as in Run 1) and RoBERTa model trained with the augmented dataset for English predictions. This was done to observe how different models in task1b perform after using the same model for classifying data in task1a.
  - **task1b** (Language-specific models): RoBERTuito model trained with the augmented dataset for Spanish predictions and RoBERTa model trained with the augmented dataset for English predictions.

- **Run 5:**
  - **task1a** (Multilingual model): XLM-Twitter trained with the original dataset for both Spanish and English predictions.
  - **task1b** (Multilingual model): XLM-Twitter model trained with the augmented dataset for both Spanish and English predictions.

Table 2 illustrates the various trained Spanish, English, and multilingual models used for task1a and task1b.

4.2. Results

The dataset used for the shared task and the description of the metrics employed to assess the results are presented in the task overview [23]. The primary metric for ranking the outcomes

Table 3
task1a's results on Spanish dataset

Model	F1-True	F1-False	F1-Macro	ICM
RoBERTuito	0.6387	0.9426	0.7906	0.1315
RoBERTuito + DA	0.6281	0.9459	0.787	0.1266
RoBERTa + DA	0.6264	0.9315	0.7789	0.1017
RoBERTuito	0.6387	0.9426	0.7906	0.1315
XLM-Twitter	0.4803	0.9109	0.6956	0.0835

was the Information Contrast Model (ICM), a metric used to provide an accurate evaluation for Multi-label Hierarchical Extreme classification [24]. Furthermore, F1-Macro was employed to evaluate the participating systems, given that both tasks involve classification.

The evaluations provided by the organizers were divided into three categories:

1. Evaluations of the predictions for the Spanish texts.
2. Evaluations of the predictions for the English texts.
3. Evaluations of the predictions for multilingual (Spanish and English texts) predictions (labeled as bilingual).

We participared in Task 1, focusing on subtasks a (propaganda identification) and b (propaganda characterization). We conducted evaluations in both Spanish and English, resulting in separate outcomes for each language, as well as a combined bilingual evaluation.

The metrics obtained for task1a evaluated on Spanish data are presented in Table 3. The best results were achieved in runs 1 and 4 (RoBERTuito), followed by run 2 (RoBERTuito + data augmentation). That is, our best-performing model for this task is RoBERTuito, trained with the original dataset. It is likely that the superiority of this model is due to the fact that it was originally trained on tweets, which are also the type of texts in the competition dataset. The use of data augmentation did not provide better performance. This suggests that the text variation from our data augmentation process was insufficient to improve model performance.

For the evaluation of task1a on the English dataset (see Table 4), run 2 (XLNet + augmented data) clearly yielded the best results, closely followed by run 1 with the RoBERTa model. Thus, the top models for binary classification in this competition were XLNet and RoBERTa.

In those two previous evaluations, the multilingual model XLM-Twitter trained with the original dataset yielded the lowest results. This is because the model is multilingual and trained with both Spanish and English data, resulting in lower precision in its predictions compared to language-specific models.

The models were also evaluated on both datasets (see Table 5). Best results were obtained by combining RoBERTuito and XLNet, where Spanish texts were classified using the former, and English texts were classified using the latter. This outcome was expected, as these models were the best performers in their respective languages (Spanish and English).

Observing the metrics obtained for each class we can see better metrics for some classes compared to others. In task1a the *true* label shows lower metrics than the *false* label, which is due to the imbalance in the training data. We attempted to balance the dataset by generating new positive text through paraphrasing, but the results did not show significant improvement.

Table 4
task1a's results on English dataset

Model	F1-True	F1-False	F1-Macro	ICM
RoBERTa + DA	0.6088	0.9255	0.7672	0.1061
XLNet + DA	0.6229	0.9289	0.7759	0.1264
BERT + DA	0.5859	0.8947	0.7403	0.0227
RoBERTa + DA	0.6088	0.9255	0.7672	0.1061
XLM-Twitter	0.4528	0.9111	0.682	-0.0573

Table 5
task1a's results on Spanish and English dataset (bilingual evaluation)

Models	F1-True	F1-False	F1-Macro	ICM
RoBERTuito - RoBERTa	0.6223	0.934	0.7782	0.1183
RoBERTuito - XLNet	0.6252	0.9374	0.7813	0.1268
RoBERTa + BERT	0.6031	0.9133	0.7582	0.0612
RoBERTuito + RoBERTa	0.6223	0.934	0.7782	0.1183
XLM-Twitter + XLM-Twitter	0.4666	0.911	0.6888	-0.0704

Table 6
task1b's results on Spanish dataset

Model	F1-1 Appeal to Commonality	F1-2 Discrediting the Opponent	F1-3 Loaded Language	F1-Macro	ICM
RoBERTa	0.2636	0.7489	0.1411	0.3845	-0.2081
RoBERTa-emoji	0.1399	0.7064	0.1672	0.3378	-0.2274
BETO	0.2541	0.6906	0.2156	0.3868	-0.3317
RoBERTuito + DA	0.2624	0.7222	0.1722	0.3856	-0.1941
XLM-Twitter	0.1852	0.4944	0.1697	0.2831	-0.5533

This may be because the new texts are too similar to the existing ones, so the model did not learn much beyond what it already knew.

Now, we will evaluate the results obtained for task1b.

The results for task1b obtained evaluating Spanish predictions are presented in Table 6. Run 4 (RoBERTuito with augmented data), achieved the highest score, closely followed by run 1 (RoBERTa without augmented data). This again demonstrates RoBERTuito's superior performance, likely due to the fact that it was trained on tweets. However, the improvement over the RoBERTa model is not meaningful, suggesting that for this multilabeling task, the model's training dataset might not be a critical factor.

The results for task1b on the English dataset are shown in Table 7. It can be observed that run 2 yielded the best results, followed by run 1 (XLNet + Data Augmentation). Run 2 used a RoBERTa-emoji multilingual model, which accounts for emojis, thereby providing better predictions. Consequently, none of the English prediction models for this task outperformed

Table 7
task1b's results on English dataset

Model	F1-1 Appeal to Commonality	F1-2 Discrediting the Opponent	F1-3 Loaded Language	F1-Macro	ICM
XLNet + DA	0.3237	0.5671	0.4352	0.442	-0.2945
RoBERTa-emoji	0.3682	0.5443	0.476	0.4628	-0.2395
RoBERTa + DA	0.313	0.5719	0.2227	0.3692	-0.4265
BERT + DA	0.274	0.5317	0.4442	0.4166	-0.3827
XLM-Twitter	0.2692	0.3864	0.3842	0.3466	-0.493

Table 8
task1b's results on Spanish and English dataset (bilingual evaluation)

Models	F1-1 Appeal to Commonality	F1-2 Discrediting the Opponent	F1-3 Loaded Language	F1-Macro	ICM
RoBERTa - XLNet	0.3067	0.6678	0.3575	0.444	-0.2421
RoBERTa-emoji - RoBERTa-emoji	0.3122	0.6383	0.3953	0.4486	-0.2205
BETO - RoBERTa	0.2939	0.6375	0.2197	0.3837	-0.3725
RoBERTuito - BERT	0.2713	0.6389	0.3906	0.4336	-0.2907
XLM-Twitter - XLM-Twitter	0.2354	0.4569	0.3088	0.3337	-0.5127

the multilingual model that incorporated emojis and was trained on the original dataset without augmentation.

For task1b, we have obtained better metrics for some classes compared to others. There is a notable difference in metrics among the three classes. In the Spanish dataset, there are almost twice as many texts for the label *Discrediting the Opponent* compared to the other two labels, which is reflected in the better metrics for that class. In the English dataset, there is an imbalance for the class *Appeal to Commonality*, which has approximately 200 fewer instances than the other two classes, resulting in the lowest F1 value. Although the classes *Discrediting the Opponent* and *Loaded Language* have a similar number of instances, the metrics for the second one are slightly lower in most cases.

In general, as metrics obtained for task1b could be affected by the predictions obtained in task1a, we decided to evaluate this by generating the predictions for task1a with the same models in runs 1 and 4. By analyzing the results obtained for task1b in runs 1 (RoBERTa & XLNet models) and 4 (RoBERTuito & BERT models), we observed that the difference between the evaluation results was minimal. Comparing their results with those obtained from other runs, we see that some models achieved significantly lower results. Therefore, we can conclude that the models used for obtaining predictions in task1a have a noticeable effect on the results obtained for task1b.

In Table 9, we present the rankings for our group's results alongside the best and worst metrics obtained in the general competition, categorized by task and language. A total of 8

Table 9
Ranking for the obtained results

	Group metrics				Best Run	Worst Run
Language	**Best Run**	**Group rank**	**Run rank**	**ICM**	**ICM**	**ICM**
			Task1a			
Spanish	1 & 4	5	12	0.1315	0.2187	-0.3371
English	2	5	14	0.1264	0.2123	-0.8796
Multilingual	2	5	13	0.1268	0.2048	-0.6176
			Task1b			
Spanish	4	4	9	-0.1941	-0.0917	-0.8013
English	2	4	9	-0.2395	0.0312	-1.4863
Multilingual	2	4	8	-0.2205	-0.0074	-1.1826

groups participated in task1a, submitting 32 runs in total, and 4 groups participated in task1b, submitting 15 runs. Additionally, since two groups did not submit their results in the correct format, they received a score of 0 for all metrics. Consequently, the worst metric excludes these groups.

Upon comparing our results with the ones obtained by the other groups, we observe that for task1a, our ICM metric is approximately 0.1 lower than the best metric across all language evaluations. This is a commendable performance, considering that only two groups achieved a value greater than 0.19. For task1b, in both English and Spanish evaluations, our ICM metric is also 0.1 lower than the best metric. However, in the bilingual evaluation for task1b, the difference from the best metric is 0.2, indicating a more significant gap and relatively poorer performance compared to other competitors.

In all our evaluations, the results we obtained are significantly better than the lowest results recorded. The lowest metric consistently appears in the run labeled *01-task-1-ensemble-highest_bilingual*.

The DIPROMATS competition was also held last year, and it is clear that the overall results in 2024 have improved. For task1a, the highest score in the previous year was achieved in the English dataset, with a F1-Macro of 0.809 and an ICM of 0.2013. This year, the best metrics are 0.813 and 0.2123, respectively. Although there is some improvement, it is not very significant compared to the previous year. However, for task1b, there is a notable improvement over the previous year's competition. The results our team obtained are 0.7906 for F1-Macro and 0.1315 for ICM, which are slightly lower than the best ones obtained in the previous year. In 2023, the best metrics for that task were obtained by the NL4IA team [2] for the English data, with a F1-Macro of 0.5591 and an ICM of 0.1778. This year, these metrics have improved to 0.6219 and 0.0312, respectively. For task1b evaluated on English dataset, the metric we obtained were 0.4628 for F1-Macro and -0.2395, which are also lower than the best ones in 2023. It should be noted that one label has been removed compared to last year's competition, making the task simpler and likely contributing to better results.

5. Conclusions

In this paper, we have proposed an approach to solving different tasks from the DIPROMATS 2024 tasks at IberLEF 2024. Our work has been focused on solving the propaganda identification and the coarse-grained propaganda classification tasks. We have showed the effectiveness of different LLMs on solving these tasks. We have tried differently structured models, mainly BERT-like, GPT-like and XLNet-like models. These models included language specific models for English and Spanish, as well as multilanguage models.

Based on our results, we observed that the best model for classifying and labeling Spanish propagandistic tweets is RoBERTuito, as it is also trained on a tweets' dataset. For English texts in task1a, the XLNet model performed best, while for task1b, the best model was RoBERTa-emoji, a multilingual model.

Our analysis also indicates that using techniques such as paraphrasing and translation to balance and augment the data does not yield significantly better results than using the original dataset.

We consider that we have obtained competent results in the task and that our approach can help in the identification of propagandistic texts, which have a great impact on audiences beliefs and perceptions.

In future work, we plan to apply other LLMs, such as LLaMA, which is known for its efficiency and ability to handle a diverse range of natural language processing tasks with high accuracy [25]. Additionally, we would like to investigate various data augmentation techniques that take into account the preservation of emoticons during the generation of new data.

References

[1] P. Moral, G. Marco, J. Gonzalo, J. Carrillo-de Albornoz, I. 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) 397–407.

[2] A. Pritzkau, Investigating propaganda considering the discursive context of utterances (2023).

[3] M. Casavantes, M. Montes-y Gómez, D. I. Hernández-Farías, L. C. González, A. Barrón-Cedeño, Propaltl at DIPROMATS: Incorporating contextual features with bert's auxiliary input for propaganda detection on tweets (2023).

[4] F. Jáñez-Martino, A. Barrón-Cedeño, Unileon-unibo at iberlef 2023 task DIPROMATS: Roberta-based models to climb up the propaganda tree in english and spanish (2023).

[5] L. Tian, X. Zhang, M. M.-H. Kim, J. Biggs, Efficient text-based propaganda detection via language model cascades (2023).

[6] J. A. García-Díaz, R. Valencia-García, Umuteam at DIPROMATS 2023: Propaganda detection in spanish and english combining linguistic features with contextual sentence embeddings (2023).

[7] V. Ahuir, L. F. Hurtado, F. García-Granada, E. Sanchis, Elirf-vrain at DIPROMATS 2023: Cross-lingual data augmentation for propaganda detection (2023).

[8] Sai Vamsi Alisetti, Paraphrase-Generator, 2020. URL: https://github.com/Vamsi995/Paraphrase-Generator.

[9] J. Tiedemann, S. Thottingal, OPUS-MT – Building open translation services for the World, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.

[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (2019).

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018).

[13] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: PML4DC at ICLR 2020, 2020.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019).

[15] 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 (2019).

[16] A. Gutiérrez Fandiño, J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, M. Villegas, MarIA: Spanish Language Models, Procesamiento del Lenguaje Natura 68 (2022).

[17] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[18] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks (2021).

[19] M. García-Vega, M. Díaz-Galiano, M. García-Cumbreras, F. Del Arco, A. Montejo-Ráez, S. Jiménez-Zafra, E. Martínez Cámara, C. Aguilar, M. Cabezudo, L. Chiruzzo, Overview of TASS 2020: Introducing emotion detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020, pp. 163–170.

[20] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification (2020).

[21] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding (2019).

[22] F. Barbieri, L. E. Anke, J. Camacho-Collados, XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond (2021).

[23] P. Moral, J. Fraile, G. Marco, A. Peñas, J. Gonzalo, Overview of DIPROMATS 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 73 (2024).

[24] E. Amigo, 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.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models (2023), arXiv preprint arXiv:2302.13971 (2023).