VICTOR VECTORS @ DIPROMATS 2024: Propaganda Detection with LLM Paraphrasing and Machine Translation

Miguel Fernández[1,3,†], Maximiliano Ojeda[1,3,†], Lilly Guevara[2,4,†], Diego Varela[2,4,†], Marcelo Mendoza[1,2,3,*,†] and Alberto Barrón-Cedeño[5,†]

[1]Pontificia Universidad Católica de Chile, Vicuña Mackenna 6840, Santiago, Chile
[2]National Center of Artificial Intelligence (CENIA), Vicuña Mackenna 6840, Santiago, Chile
[3]Millennium Institute on Foundational Research on Data (IMFD), Vicuña Mackenna 6840, Santiago, Chile
[4]Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile
[5]Università di Bologna, Corso della Repubblica 136, Forlì, Italy

Abstract

Identifying propaganda in social media posts is an important task that can help to better understand the strategies applied by policy makers and stake holders when trying to convey their message to the general public. We describe our participation in DIPROMATS 2024 Task 1 on the automated detection and characterization of propaganda techniques and narratives from diplomats of major powers. We show an efficient way to utilize Large Language Models (LLMs) to paraphrase a sample of the training instances, to balance the class distribution in the datasets provided by the shared task. Our submission ranked 1st in Subtask-1a in English (ICM score of 0.2123) and 1st in the bilingual evaluation (ICM score of 0.2048). We also achieved top-3 rankings in Spanish and subtasks 1b and 1c.

Keywords

Propaganda detection, LLMs, Paraphrasing, Data augmentation, Unbalanced data

1. Introduction

In this study, we explore various text-based methods to address Task 1 of the DIPROMATS 2024 shared task on Automated Detection and Characterization of Propaganda Techniques and Narratives from Diplomats of Major Powers [1], part of IberLEF 2024 [2]. DIPROMATS focuses on advancing research in developing Natural Language Processing (NLP) tools to detect propaganda on social media. Our approach is based on the use of LLMs in combination with transformer encoders to identify propaganda solely from the textual content of the social media posts.

Our team centers on Task 1 —both in English and in Spanish. Task 1 challenged the participants to develop models for "Propaganda Identification and Characterization". It is divided into three subtasks. Subtask 1a focuses on propaganda identification. It involves determining whether a tweet uses propaganda techniques, framing it as a binary classification problem. Subtask 1b involves a coarse-grained approach to propaganda detection. Systems must categorize each tweet into one of four predefined classes: Not propagandistic, Appeal to commonality, Discrediting the opponent, or Loaded language. This is a multiclass, multilabel classification task. Subtask 1c deals with fine-grained propaganda detection. In this subtask, systems are required to classify messages based on finer-grained propaganda techniques. The classification includes one negative class and seven positive classes: Flag Waving, Ad Populum/Ad antiquitatem, Name Calling/Labelling, Undiplomatic Assertiveness/Whataboutism, Appeal to Fear, Doubt, and Loaded Language.

We focused on addressing the imbalances in the datasets provided for the competition, specifically in the annotations of subtasks 1b and 1c. Our initial evaluation indicated that certain classes presented

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
†These authors contributed equally.
mafernandez17@uc.cl (M. Fernández); muojeda@uc.cl (M. Ojeda); lilly.guevara@usm.cl (L. Guevara); diego.varelaf@sansano.usm.cl (D. Varela); marcelo.mendoza@uc.cl (M. Mendoza); a.barron@unibo.it (A. Barrón-Cedeño)

more challenges than others. To mitigate this, we employed two data augmentation strategies aimed at improving these classes. The first strategy involved translating text samples between Spanish and English, thereby generating new instances in each language's dataset. The second approach utilized paraphrasing through LLM prompts, creating cross-lingual transference of samples between Spanish and English. This method significantly enhanced our performance in subtask 1a, leading our team to achieve first place. However, its effectiveness diminished with the increased specificity of subtasks 1b and 1c. For these, the translation-based data augmentation proved more successful, particularly when transferring knowledge from Spanish to English, as opposed to the other way around.

The rest of the paper is distributed as follows. Section 2 overviews background knowledge on the task. Section 3 describes the provided datasets. Section 4 introduces our methodology, whereas Section 5 describes our models. Section 6 presents our experiments and discusses the obtained results. Section 7 closes with conclusions and future work.

2. Background

Far from reducing the impact of politics becoming increasingly media-driven, the rise of social networks has expanded the reach of disinformation and propaganda [3]. In traditional mass media environments, propaganda was often seen as a top-down process, which Ellul referred to as "political propaganda" [4]. This contrasts with "cultural propaganda," which spreads horizontally and disguises its propagandistic nature within what appears to be organic conversations. The lines between these types of propaganda blur in the digital age, where both forms can achieve widespread diffusion.

Political propaganda can be subtly introduced through networks of ordinary-seeming users. Research confirms that repetition and stereotyped language play significant roles in radicalizing views and fostering fertile environments for disinformation campaigns, enhanced by the unique features of digital media [5]. This phenomenon has been observed in multiple geopolitical events, including the 2016 US presidential election [6], the 2019 UK general election [7], the White Helmets campaign [8], Russian trolls' interventions in the US [9], the social unrest in Chile [10], and the Russian invasion of Ukraine [11]. In all these cases, radicalization was central [12, 13], often linked to coordinated propaganda campaigns [14]. Some studies suggest that a decrease in the threshold for verification [15], leading to increased susceptibility to propaganda, is influenced not only by political bias but also by the readability typical of propagandistic texts [16]. In this context, fake news, media bias, and propaganda are all components of a disinformation ecosystem [17].

Propaganda promotes political polarization, and opinion leaders such as politicians and journalists play a crucial role in this two-step flow of information [18]. They set the agenda by either confirming or challenging facts, and technology can be used to either heighten the visibility of these campaigns [19] or drown them in a flood of noise [20]. In this ecosystem, the context of social media alters how information is discerned [21].

The advent of new language technologies such as Chat-GPT [22] and various distilled versions of open-source LLM [23] has inadvertently provided malevolent entities with powerful tools for disinformation and propaganda [24]. These technologies enable the creation of human-like content, potentially ushering in a new era of actors and phenomena on social networks that could intensify the effects of disinformation and propaganda campaigns [14]. The extent of these coordinated efforts is currently unpredictable and poses significant threats not only to the integrity of the information ecosystem but also to democratic processes [25].

Ironically, these same technologies offer us opportunities to mitigate the negative impact of disinformation and propaganda on social networks. We aim to develop new language technologies, underpinned by LLMs [26, 22], to identify propaganda techniques and verify their usage by news outlets and social media influencers in real-time. Our approach includes deploying detectors specifically designed to recognize propaganda disseminated through news and rumors on social platforms. This initiative is intended to augment the work of fact-checkers, who primarily focus on confirming the accuracy of information [10]. While their efforts are crucial and deserve greater support, they must be comple-

mented by strategies to detect propaganda, given that not all propaganda is inherently false [8]. Often, truthful information can be manipulated to produce outcomes favorable to specific groups [17]. Our objective is to combat these orchestrated campaigns by increasing awareness, exposing them, and providing tools that help a broad audience —from everyday social media users to public figures and media professionals— understand the true nature of the information they encounter. We begin this effort by evaluating some of our methodologies at DIPROMATS 2024, a challenge that asks participants to detect the use of propaganda in texts. These texts, spread across social networks by prominent politicians, are presented in various languages, including English and Spanish.

3. Dataset

The task corpus provided consists of tweets in both Spanish and English posted by diplomats representing four international actors: China, Russia, the United States, and the European Union. These tweets come from various diplomatic sources, including government accounts, embassies, ambassadors, consuls, and missions. For Task 1, which focuses on the identification and characterization of propaganda, the corpus includes two annotated datasets: one with tweets in English and another in Spanish. These tweets were collected using the Twitter API for Academic Research and were posted between January 1, 2020, and March 11, 2021, the latter date marking the first anniversary of the COVID-19 pandemic declaration.

The Spanish dataset comprises 9,591 tweets posted by 135 diplomatic authorities, distributed as follows: 2,997 tweets from 25 Chinese authorities, 1,391 from 22 Russian authorities, 2,465 from 48 European Union authorities, and 2,738 from 40 US authorities. The English dataset contains 12,012 tweets from 619 authorities, with the distribution being: 3,022 tweets from 106 Chinese authorities, 2,690 from 114 Russian authorities, 2,916 from 186 European Union authorities, and 3,114 from 216 U.S. authorities.

The datasets are divided into two temporal partitions for the task. The older 70% subset is used for training, while the newer 30% subset is reserved for testing. The test data is kept private to prevent overfitting in post-campaign experiments.

The tweets are annotated for three subtasks. A significant aspect of the dataset is its multilabel nature: a tweet can simultaneously belong to one or more classes within each subtask. Table 1 illustrates the class distribution in the DIPROMATS 2024 dataset. The figures show a significant imbalance in the number of instances per class, which poses a challenge to training models. Furthermore, given that the annotation scheme is multilabel, there are interdependencies among the classes. To illustrate the relationships between classes based on intersectionality, Figures 1 and 2 depict the intersectionality for subtask 1c, for English and Spanish, respectively.

4. Methodology

As shown in Table 1, the class distribution across the three subtasks display a significant imbalance. To counter this issue, we evaluated various data augmentation techniques to improve the balance between classes.

One such technique involved the creation of examples through paraphrasing. This method was applied to augment the English training set. We implemented it with consideration for the Subtask 1c and used the GPT-3.5 model. The prompt used to generate examples was as follows:

Language	Subtask	Label	#
English	1a	No propaganda	6,437
		Propaganda	1,971
Spanish	1a	No propaganda	4,924
		Propaganda	1,196
English	1b	No propaganda	6,437
		1 appeal to commonality	607
		2 discrediting the opponent	910
		3 loaded language	913
Spanish	1c	No propaganda	6,439
		1 appeal to commonality - ad populum	72
		1 appeal to commonality - flag waving	545
		2 discrediting the opponent - doubt	74
		2 discrediting the opponent - appeal to fear (destructive)	57
		2 discrediting the opponent - name calling	213
		2 discrediting the opponent - undiplomatic assertiveness/whataboutism	681
		3 loaded language	913
Spanish	1b	No propaganda	4,924
		1 appeal to commonality	291
		2 discrediting the opponent	671
		3 loaded language	389
Spanish	1c	No propaganda	4,924
		1 appeal to commonality - ad populum	59
		1 appeal to commonality - flag waving	234
		2 discrediting the opponent - doubt	27
		2 discrediting the opponent - appeal to fear (destructive)	61
		2 discrediting the opponent - name calling	90
		2 discrediting the opponent - undiplomatic assertiveness/whataboutism	569
		3 loaded language	389

Table 1
Human annotations across data partitions and subtasks.

Prompt

instruction = f""" Your job is to paraphrase the phrase that I will give you, which contains a propaganda technique in Spanish. The objective is to maintain the use of that technique in the paraphrased result. Now I will show you the name of the technique, the definition and an example, so you can understand it better. """

out = f""" The expected output is the result of paraphrasing the following sentence: """

def paraphrase(technique, definition, example, tweet, temperature=0.7):

```
    try:
      response = openai.chat.completions.create(model = "gpt-3.5-turbo",
        messages = [{"role": "user",
                     "content": instruction + f'''"{technique}", "
                     {definition}"'''+ f'''"{example}"'''+ out
                     + f'''"{tweet}"'''}],
        temperature = temperature)
        answer = response.choices[0].message.content

        return answer
```

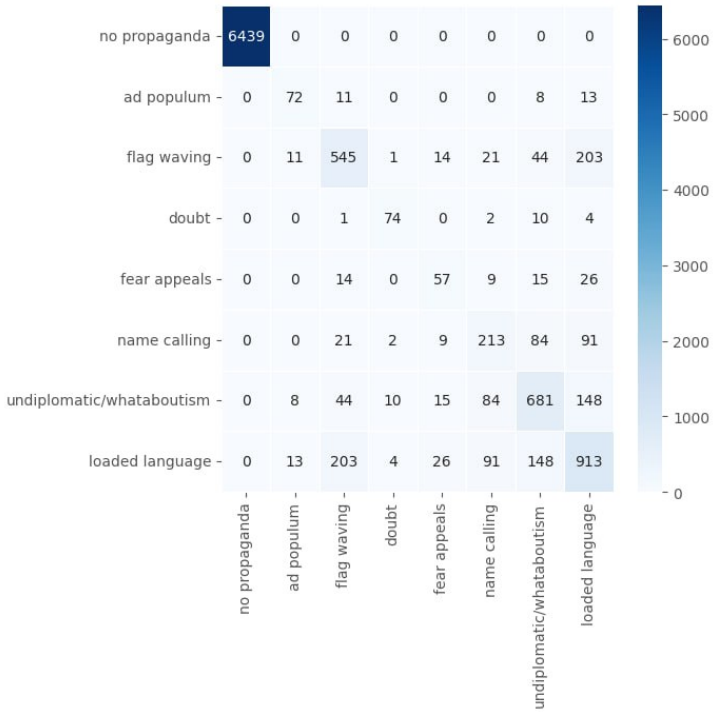

Figure 1: Intersectionality of the classes for Subtask 1c (English).

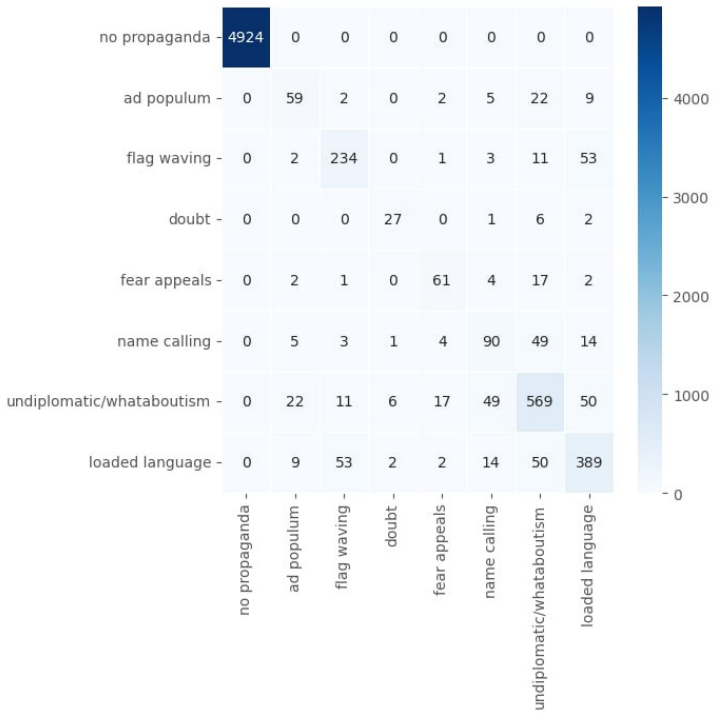

Figure 2: Intersectionality of the classes for Subtask 1c (Spanish).

The "paraphrase" function provides the name, definition, and example of the desired technique. These definitions were obtained from the DIPROMATS 2024 website.[1] Consequently, the hard prompt establishes a one-shot learning task, where the style of propaganda from the example in Spanish must be transferred to the tweet in English.

[1] https://sites.google.com/view/dipromats2024

Table 2
Evaluation the impact of various combinations of selected classes for data augmentation. The seven possible class combinations, derived from the three most challenging classes, were assessed across three different augmentation strategies. Notation: FA: appeal to fear, FW: Flag Waving, Ad Populum: AP.

Exps	RoBERTa + transl. ES-EN	RoBERTa-S + transl. ES-EN	RoBERTa + paraphrasing EN
# 1	FA	**FA**	FA
# 2	FW	FW	FW
# 3	**AP**	AP	AP
# 4	FA - FW	FA - FW	FA - FW
# 5	FA - AP	FA - AP	**FA - AP**
# 6	FW - AP	FW - AP	FW - AP
# 7	FA - FW - AP	FA - FW - AP	FA - FW - AP

A second technique for creating examples for minority classes involves using machine translation. To augment the dataset, we translate instances from the English dataset into Spanish, and vice versa. We implemented this data augmentation strategy using the Helsinki-NLP/opus-mt-en-es and Helsinki-NLP/opus-mt-es-en models (see https://huggingface.co/Helsinki-NLP).

We evaluated the impact of the two augmenting strategies —paraphrasing and translation— on the performance of the classifier. In this preliminary evaluation, we used a RoBERTa-base encoder [27]. Training a linear classifier on the base of the text embeddings provided by RoBERTa-base, we were able to classify texts into one of the seven categories of Subtask 1c. This initial assessment allowed us to identify that the categories where this classifier performed poorly, according to the F1 score at the macro level for both languages, were Appeal to Appeal, Flag Waving, and Ad-Populum. Building on this insight, we assessed the performance impact of applying data augmentation techniques to these classes. Since the classes may have interdependencies, we found it relevant to evaluate different combinations of categories for data augmentation. Considering the three classes involved, we derived seven types of class combinations, as shown in Table 2.

Regarding the augmentation of examples per class, we considered increasing them by 50%, based on the class distribution of the original training. To carry out the evaluation, we partitioned the training set in a 90/10 ratio, reserving 10% of the validation for testing. According to this partitioning strategy, the number of examples created for each class was: appeal to fear: 23, Flag Waving: 40, and Ad Populum: 35.

As shown in Table 2, the initial evaluation was conducted on two different augmentation strategies: one based on ES-EN translation and another on EN paraphrasing. We used RoBERTa-base as the text encoder for both strategies. Additionally, we evaluated a variant of RoBERTa fine-tuned on data from SemEVAL 2020, task 11 [28], which focuses on the detection of propaganda techniques in news articles. The task used for the RoBERTa fine-tuning was multiclass classification. Based on this encoder tuned with SemEVAL data, denoted by RoBERTa-S, we assessed the effectiveness of data augmentation based on ES-EN translation.

For each of the evaluated strategies, it was found that a specific type of augmented class yielded the most benefits. In the case of RoBERTa + ES-EN translation, the best results were achieved by augmenting the Ad Populum class. The evaluation of RoBERTa-S indicated that the best results are obtained by augmenting the Appeal to Fear class. Finally, when using paraphrasing, the best results were achieved by augmenting both the Ad Populum and appeal to fear classes. This was the only case where a benefit was observed from augmenting more than one class.

Working with the Spanish data provided by DIPROMATS 2024, we replicated the methodology used in English. Specifically, for the three classes with the lowest performance in Subtask 1c, we assessed the impact on classifier performance of augmenting combinations of these three classes. In this case, only the augmentation approach based on translation from English to Spanish was evaluated. Regarding the text encoders, we tested three different models: RoBERTa [27], BETO [29], and multilingual BERT [30]. Among these, BETO yielded the best results. Consequently, BETO was used for all subtasks of task 1

that involved examples in Spanish.

A key aspect we focused on was the process to select examples for augmentation, either through translation or throush paraphrasing. We intended to avoid collateral effects due to the augmentation of examples, based on interdependencies between classes. This is crucial because creating a new example in one class could potentially intersect with another class, leading to a noisy increase in both, thereby exacerbating the imbalance rather than reducing it. To prevent such collateral effects of data augmentation, we developed a technique for selecting examples. The idea was implemented through the following procedure. Using SentenceBERT [31] with the 'all-MiniLM-L6-v2' model, we created embeddings for all tweets belonging to the three most difficult classes. Subsequently, a selection process was carried out following these steps:

1. The cosine similarity between the embeddings of the three classes was calculated. For example, for the tweets of "appeal to fear," the cosine similarity between "appeal to fear" vs. "Flag waving" and "appeal to fear" vs. "Ad populum" was calculated.
2. For each class pair, the tweets were sorted according to cosine similarity to identify the pairs that were most distant from each other.
3. Considering a pivot class, tweets from this class that were most distant from the other two were selected. For example, taking "appeal to fear" as the pivot, tweets from this class that had the least semantic similarity against all the tweets from the "Flag Waving" and "Ad populum" classes were chosen.

In this way, the tweets used to augment a particular class were those most distant from the other two classes. The amount of instances selected was sufficient to reach the pre-established quantity before this selection process began. For instance, the 23 tweets of "appeal to fear" that were most distant from the other two were selected. This process was repeated for each of the three most challenging classes in DIPROMATS 2024, Subtask 1c.

5. Models

Unlike the data exploration stage where we used RoBERTa-base as a text encoder, for the training stage of competitive models for English data, we based our work on the RoBERTa-Large architecture. Regarding hyperparameters, and based on the hyperparameter exploration conducted by Garcia-Diaz and Valencia-Garcia [32] as part of DIPROMATS 2023, we replicated the parameters to train a classifier based on RoBERTa-Large. This configuration includes 5 epochs, a batch size of 32, a learning rate of 2e-5, and a weight decay of 0.18. These same parameters were used for the models trained for Subtasks 1a, 1b, and 1c.

For English, the following models were trained:

1. **Model 1: RoBERTa-Large + data augmentation based on translation from Spanish to English in the Ad Populum class**. The pre-trained RoBERTa text encoder was fine-tuned on the DIPROMATS 2024 dataset, which had been augmented by translating examples from Spanish to English in the Ad Populum class of Subtask 1c.
2. **Model 2: RoBERTa-Large + fine-tuning on SemEVAL 2020 task 11 + data augmentation based on translation from Spanish to English in the Ad Populum class**. The RoBERTa-Large model was trained using the SemEVAL 2020 task 11 dataset. This competition focused on identifying propaganda techniques in news articles using a fine-grained detection approach. To adapt this dataset to a multi-label classification approach, each article was transformed into a group of sentences by extracting each line, resulting in one sentence per row. This dataset was used to fine-tune RoBERTa-Large for the multi-label task, considering the 14 predefined classes. Subsequently, the fine-tuned model was further fine-tuned on the DIPROMATS 2024 data for each of the three subtasks, incorporating the data augmentation approach focused on the appeal to fear class through Spanish to English translation.

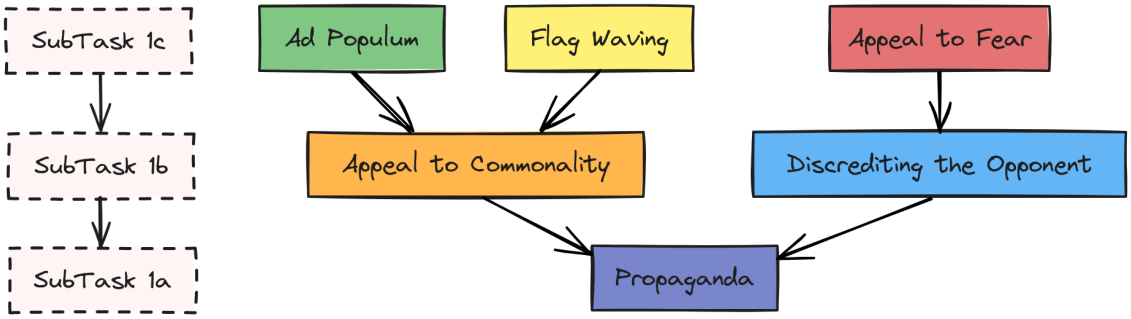

Figure 3: Schema for transferring examples obtained through data augmentation across subtasks.

Table 3
Preliminary evaluation based on F1-macro.

Model	Subtask 1a	Subtask 1b	Subtask 1c
Model 1 (English)	0.82	0.72	0.67
Model 2 (English)	0.82	0.72	0.67
Model 3 (English)	0.83	0.72	0.68
Model 4 (Spanish)	0.87	0.67	0.58

3. **Model 3: RoBERTa-Large + data augmentation based on paraphrasing of examples from the Ad Populum and appeal to fear classes for subtask 1c**. This third model involves data augmentation through paraphrasing for two minority classes: Ad Populum and appeal to fear. In this case, the predictions made for subtask 1c were used to infer the labels for subtasks 1a and 1b.

For Spanish, we used a text encoder based on BETO (**Model 4**). This encoder was trained on a version of the DIPROMATS 2024 dataset augmented in the Flag Waving and appeal to fear classes. In this case, the technique based on translation from English to Spanish was used.

To transfer the augmented examples from Subtask 1c to the other subtasks, we followed this procedure. First, we used the class containment relationships defined by DIPROMATS 2024 in the definition of strategies. Specifically, the augmented examples in Ad Populum and Flag Waving correspond to a class that contains them, known as Appeal to Commonality. Consequently, all the augmented examples in the Ad Populum and Flag Waving classes contributed to increasing the Appeal to Commonality class. According to the definition of strategies by DIPROMATS 2024, class Appeal to Fear is contained within the techniques grouped under Discrediting the Opponent. As a result, the augmented examples in the Appeal to Fear class contributed to increasing the examples of Discrediting the Opponent. Finally, all the augmented examples in the three most challenging classes of Subtask 1c contributed to increasing the Propaganda class in Subtask 1a (binary classification). Figure 3 illustrates the schema for transferring examples obtained through data augmentation across subtasks. It is noted that all the generated examples ultimately contribute to achieving a diverse data augmentation within the propaganda class of Subtask 1a.

6. Experiments

6.1. Preliminary evaluation

We began the experiments by conducting an internal evaluation of the proposed models. This evaluation used a test partition derived from the training dataset provided in the competition (10% of the training set). Table 3 summarizes the results obtained using the F1-macro score for each subtask.

We disaggregate the results by class to understand the difficulty of subtasks 1b and 1c. Figure 4 displays the disaggregated performance in subtask 1b, while Figure 5 illustrates the performance in

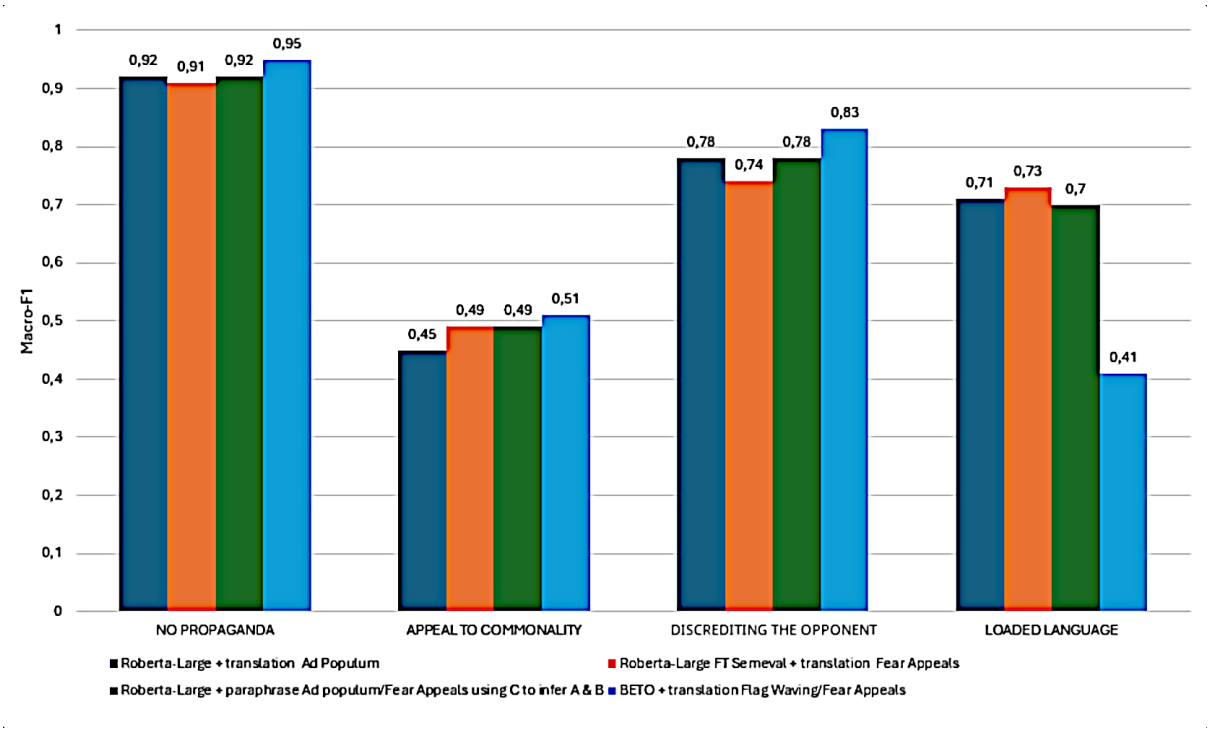

Figure 4: Preliminary evaluation based on F1-macro on subtask 1b.

subtask 1c.

Figure 4 illustrates that for all four models. The simplest task is detecting no propaganda, where the macro F1 scores exceed the 0.9 threshold. Model 4 (BETO) achieves the highest performance for this class. On the other hand, the most challenging propaganda technique for these models under Subtask 1b, is 'Appeal to Commonality.' Only one model (Model 4) surpasses the 0.5 threshold in macro F1 score for this class. For the 'Discrediting the Opponent' category, all models perform around 0.78 macro-F1 score, suggesting it is a relatively less challenging class. Here Model 4 (BETO) shows superior performance with a 0.83 score. However, in the 'Loaded Language' category, the performance of Model 4 significantly declines to a mere 0.41 macro F1 score, whereas Models 1-3 maintain a performance above 0.7. This discrepancy in the 'Loaded Language' class results in the first three models outperforming Model 4 by 5 percentage points in the overall F1 metric at the macro level (see Table 3).

Our experimentation reveals the following insights regarding Subtask 1c. As Figure 5 shows, detecting non-propaganda remains the simplest challenge for all models, each surpassing the 0.9 threshold in macro F1 score. However, when analyzing the seven propaganda classes within Subtask 1c, two techniques pose significant difficulties for the models: Ad Populum and Flag Waving. These classes have consistently presented the most challenges in our primary evaluation, along with appeal to fear. The performance across models varies considerably for these techniques. Model 1 performs best on Ad Populum, achieving a macro F1 score of 0.57, whereas Model 3 excels in Flag Waving with a score of 0.48. Notably, Model 4 (BETO) performs poorly on Ad Populum, securing only a 0.17 in F1 score. Appeal to fear also show varied results, with Models 1-3 significantly outperforming Model 4 by over 20 percentage points; Model 4 achieves a mere 0.4 in macro F1 score, whereas Models 1-3 exceed 0.6.

Furthermore, some models exhibit strong performances over specific classes. For instance, Model 2 stands out in Name Calling, while Model 3 shows strength in Doubt. However, none of the four models clearly dominate in Subtask 1c overall. Model 3 slightly leads with a macro F1 score of 0.68, whereas Model 4 lags behind with the lowest overall performance, achieving only 0.58 in macro F1 score (see Table 3).

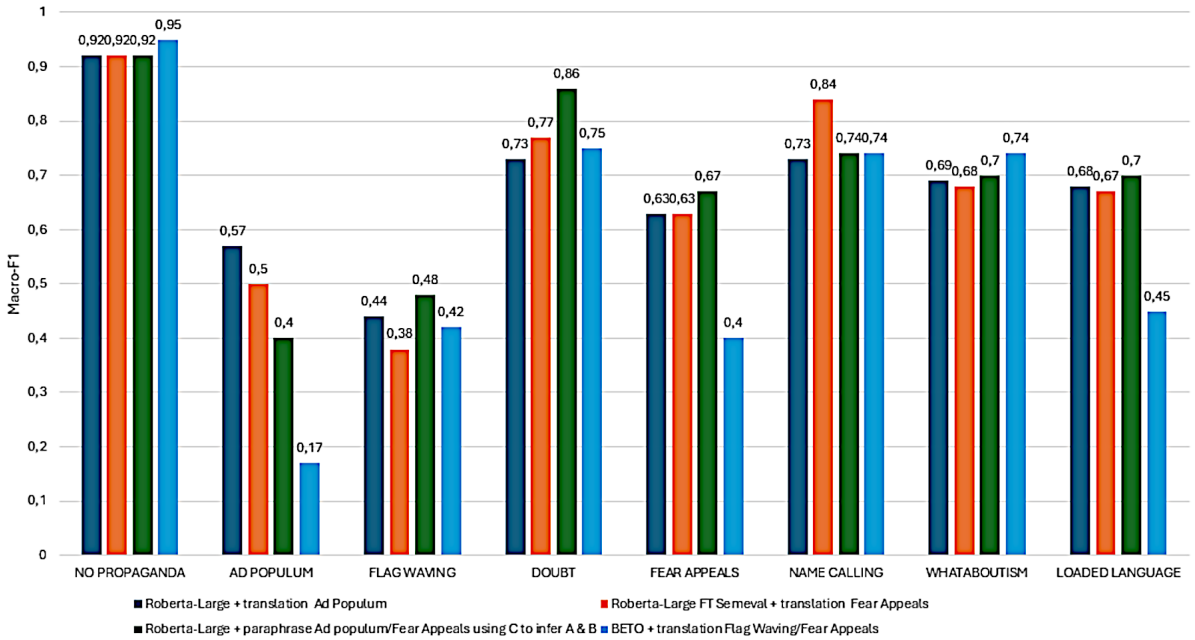

Figure 5: Preliminary evaluation based on F1-macro on subtask 1c.

6.2. Official results at DIPROMATS 2024

We submitted five runs for the official evaluation at the competition. Runs 3 and 4 were experiments using BETO, enhanced with Spanish examples augmented through paraphrasing techniques. We call this setting **Model 5**. Since subtasks 1a, 1b, and 1c included examples in both English and Spanish, the runs were always generated using two models, one for each language. The language of each example was identified using Spacy's FastLang detector. Spanish examples were consistently analyzed using Model 4 (BETO), as detailed in the previous section. Conversely, English examples were processed by Models 1 to 3. This strategy resulted in five runs. Specifically, each run corresponds to the following configurations:

Run 1 Model 1 (English) + Model 4 (Spanish).

Run 2 Model 2 (English) + Model 4 (Spanish).

Run 3 Model 1 (English) + Model 5 (Spanish)

Run 4 Model 2 (English) + Model 5 (Spanish)

Run 5 Model 3 (English) + Model 4 (Spanish).

The competition employed F1 and ICM (Information Contrast Model) as metrics, with ICM serving as the official metric for ranking the teams. For each task, the competition reported outcomes based on performance in English, Spanish, and both languages (bilingual). Table 4 displays the official results obtained by our best three runs in each of these scenarios.

Based on our best run in each scenario, the official results indicate that we achieved first place in two out of nine evaluation scenarios, second place in four, and third place in the remaining three scenarios. Overall, our best performances were observed in Task 1a, with competitive results in subtasks 1b and 1c. The most notable model was from run 5, which excelled in subtask 1a. For subtasks 1b and 1c, runs 1 and 2 were highlighted. This suggests that while Model 3 was better for binary classification in English (paraphrasing), the fine-grained tasks (subtasks 1b and 1c) benefited more from Models 1 and 2, both utilizing data augmentation with translation. These findings indicate that the effectiveness of

Table 4
Official results at DIPROMATS 2024, including the official ranks at the shared task leaderboard.

Scenario	Rank	Run	ICM	F1-true	F1-false	F1-macro
Task 1a - bilingual	1	5	0.204	0.691	0.943	0.816
	2	2	0.203	0.689	0.942	0.816
	3	1	0.202	0.688	0.942	0.815
Task 1a - English	1	5	0.212	0.691	0.935	0.813
	2	1	0.209	0.689	0.935	0.812
	4	2	0.206	0.687	0.934	0.811
Task 1a - Spanish	3	5	0.195	0.691	0.951	0.821
	4	4	0.195	0.691	0.951	0.821
	5	3	0.195	0.691	0.951	0.821
Task 1b - bilingual	2	1	-0.042	-	-	0.586
	3	5	-0.055	-	-	0.581
	4	3	-0.059	-	-	0.583
Task 1b - English	2	1	0.009	-	-	0.628
	4	5	-0.024	-	-	0.604
	5	3	-0.026	-	-	0.614
Task 1b - Spanish	3	2	-0.125	-	-	0.464
	4	3	-0.125	-	-	0.464
	5	5	-0.125	-	-	0.464
Task 1c - bilingual	2	1	-0.114	-	-	0.479
	3	2	-0.115	-	-	0.521
	4	3	-0.122	-	-	0.508
Task 1c - English	2	1	-0.041	-	-	0.448
	3	2	-0.042	-	-	0.546
	5	3	-0.061	-	-	0.545
Task 1c - Spanish	3	1	-0.216	-	-	0.393
	4	2	-0.216	-	-	0.393
	5	3	-0.216	-	-	0.393

data augmentation techniques varies depending on the task granularity. In scenarios that included only Spanish examples, the best results were consistently achieved by runs 5, 2, and 1, outperforming runs 3 and 4. This shows that Model 5 (BETO + paraphrasing) was not effective. Model 4 (BETO + translation) always performed better. In summary, while paraphrasing was useful for subtask 1a, translation proved more beneficial for tasks requiring finer granularity (subtasks 1b and 1c).

7. Conclusions

In this paper, we present the results achieved by our team at DIPROMATS 2024. A significant aspect of our experimentation involved efforts to enhance the datasets provided by the competition, which exhibited severe imbalances, particularly in the annotations for subtasks 1b and 1c. An initial assessment revealed that some classes were significantly more challenging than others. To address this, we focused on improving these classes using two data augmentation techniques: one involved translating examples from Spanish to English and vice versa, creating new samples in the target language's example partition. The second strategy used paraphrasing with LLM prompts, employing examples in Spanish or English to generate new samples in the target language (cross-lingual transference through prompts). This approach yielded excellent results in subtask 1a, securing first place in this subtask for our team. However, it proved less effective as the task granularity increased. For subtasks 1b and 1c, the best

results were achieved through data augmentation based on translation, demonstrating that transferring knowledge from Spanish to English examples was more effective than the reverse.

The task presents several challenges. First, the annotations in the examples are interdependent. Properly using these interdependencies in augmentation strategies is challenging, as it requires the creation of synthetic examples without affecting the interclass relationships—an aspect we have earmarked for future work. Secondly, given the class imbalance in the provided data, techniques tend to show varied performance across classes. Improving the dataset construction phase remains a significant challenge. Finally, incorporating context is crucial for these models, as the evaluated tasks primarily involve sentence-level classification. Exploring the use of external sources or applying these techniques in conversational contexts is a promising avenue for future research.

Acknowledgments

L. Guevara, D. Varela, and M. Mendoza are supported by the National Center for Artificial Intelligence (CENIA FB210017, Basal ANID). M. Fernández and M. Ojeda acknowledge support from the Millennium Institute for Foundational Research on Data (IMFD ANID - Millennium Science Initiative Program - Code ICN17_002). M. Mendoza was supported by ANID Fondecyt grant 1241462. The funders played no role in the design of this study.

References

[1] 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).

[2] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[3] D. Lazer, M. Baum, Y. Benkler, A. Berinsky, K. Greenhill, F. Menczer, M. Metzger, The science of fake news, Science 359 (2018) 1094–96. URL: https://doi.org/10.1126/science.aao2998. doi:10.1126/science.aao2998.

[4] J. Ellul, Propaganda: The Formation of Men's Attitudes, 1st ed., Vintage, 1973.

[5] S. C. Woolley, P. Howard, Computational propaganda worldwide: Executive summary, 2017.

[6] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election, Science 363 (2019) 374–78. URL: https://doi.org/10.1126/science.aau2706. doi:10.1126/science.aau2706.

[7] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, M. Tesconi, Coordinated behavior on social media in 2019 UK general election, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 15, 2021, pp. 443–454.

[8] D. Pacheco, A. Flammini, F. Menczer, Unveiling coordinated groups behind White Helmets disinformation, in: Companion Proceedings of the Web Conference 2020, 2020, pp. 611–616. URL: https://doi.org/10.1145/3366424.3385775.

[9] L. Luceri, S. Giordano, E. Ferrara, Detecting troll behavior via inverse reinforcement learning: A case study of Russian trolls in the 2016 us election, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 417–27. URL: https://doi.org/10.1609/icwsm.v14i1.7311. doi:10.1609/icwsm.v14i1.7311.

[10] M. Mendoza, S. Valenzuela, E. Núñez-Mussa, F. Padilla, E. Providel, S. Campos, R. Bassi, A. Riquelme, V. Aldana, C. López, A study on information disorders on social networks during the Chilean social outbreak and COVID-19 pandemic, Applied Sciences 13 (2023) 5347. URL: https://doi.org/10.3390/app13095347.

[11] F. Pierri, L. Luceri, N. Jindal, E. Ferrara, Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine, in: Proceedings of the 15th ACM Web Science Conference, 2023, pp. 65–74.

[12] M. C. Bugueño, M. Mendoza, Learning to detect online harassment on Twitter with the Transformer, in: Machine Learning and Knowledge Discovery in Databases - International Workshops of ECML PKDD, Würzburg, Germany, September 16-20, Proceedings, Part II, volume 1168 of *Communications in Computer and Information Science*, Springer, 2019, pp. 298–306. URL: https://doi.org/10.1007/978-3-030-43887-6_23.

[13] M. Mendoza, D. Parra, Á. Soto, GENE: graph generation conditioned on named entities for polarity and controversy detection in social media, Information Processing & Management 57 (2020) 102366. URL: https://doi.org/10.1016/j.ipm.2020.102366.

[14] M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, P. Zola, Coordinated inauthentic behavior and information spreading on twitter, Decision Support Systems 160 (2022) 113819. URL: https://doi.org/10.1016/j.dss.2022.113819.

[15] G. Pennycook, D. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning, Cognition 188 (2019).

[16] L. Fazio, D. Rand, G. Pennycook, Repetition increases perceived truth equally for plausible and implausible statements, Psychonomic Bulletin & Review 26 (2019) 1705–1710.

[17] P. Nakov, G. Da San Martino, Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, Association for Computational Linguistics, Online, 2020, pp. 7–19.

[18] E. Katz, P. F. Lazarsfeld, Personal Influence, The part played by people in the flow of mass communications, Transaction Publishers, 1964.

[19] D. Effron, B. Helgason, The moral psychology of misinformation: Why we excuse dishonesty in a post-truth world, Current Opinion in Psychology 47 (2022) 101375. URL: https://doi.org/10.1016/j.copsyc.2022.101375. doi:10.1016/j.copsyc.2022.101375.

[20] E. Treré, The dark side of digital politics: Understanding the algorithmic manufacturing of consent and the hindering of online dissidence, IDS Bulletin 47 (2016).

[21] Z. Epstein, N. Sirlin, A. Arechar, G. Pennycook, D. Rand, The social media context interferes with truth discernment, Science Advances 9 (2023) eabo6169. URL: https://doi.org/10.1126/sciadv.abo6169. doi:10.1126/sciadv.abo6169.

[22] OpenAI, GPT-4 Technical Report, Technical Report, 2023. URL: https://doi.org/10.48550/ARXIV.2303.08774.

[23] W. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, A survey of large language models, arXiv (2023). URL: https://doi.org/10.48550/ARXIV.2303.18223.

[24] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, P. Nakov, A survey on computational propaganda detection, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020, pp. 4826–4832. URL: https://doi.org/10.24963/ijcai.2020/672. doi:10.24963/ijcai.2020/672.

[25] K. Aslett, A. Guess, R. Bonneau, J. Nagler, J. Tucker, News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions, Science Advances 8 (2022) eabl3844. URL: https://doi.org/10.1126/sciadv.abl3844.

[26] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, Training language models to follow instructions with human feedback, Technical Report, 2022. URL: https://doi.org/10.48550/ARXIV.2203.02155.

[27] 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, CoRR abs/1907.11692 (2019).

[28] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1377–1414. URL: https://aclanthology.org/2020.semeval-1.186.

doi:10.18653/v1/2020.semeval-1.186.
[29] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, 2023.
[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), 2019, pp. 4171–4186.
[31] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990.
[32] J. A. García-Díaz, R. Valencia-García, Umuteam at DIPROMATS 2023: Propaganda detection in Spanish and English combining linguistic features with contextual sentence embeddings, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN), Jaén, Spain, September 26, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.