Bilingual Propaganda Detection in Diplomats' Tweets Using Language Models and Linguistic Features

Arkadiusz Modzelewski[1,2,*,†], Paweł Golik[†] and Adam Wierzbicki[1]

[1]Polish-Japanese Academy of Information Technology, Poland
[2]University of Padua, Italy

Abstract

Our study presents an approach to a shared task of propaganda identification and characterization at the DIPROMATS 2024 hosted by the Iberian Languages Evaluation Forum. As the DSHacker team, we participated in the propaganda detection task, which comprised three subtasks, each with varying levels of detail in identifying propaganda types. The first subtask required binary identification of propaganda in tweets authored in either English or Spanish by diplomats and authorities from major powers. The second subtask focused on a coarse-grained classification of propaganda, while the third subtask demanded a fine-grained approach to identifying specific propaganda techniques. To tackle these challenges, we fine-tuned different BERT-based pre-trained models, including the XLM-RoBERTa model, and achieved remarkable success. Our system secured first place across all language categories, including monolingual and bilingual approaches, for the second and third subtask. Moreover, we attained high rankings in the binary propaganda classification. Our research also delves into the potential of detecting propaganda using Large Language Models with a few-shot prompting approach. We conducted experiments with two GPT models, including the recently released GPT-4o by OpenAI. Furthermore, we investigated the effectiveness of linguistic features and traditional machine learning models in propaganda detection. Overall, our study highlights our system's exceptional performance and provides valuable insights into the capabilities of modern language models and machine learning techniques in identifying propaganda.

Keywords

Propaganda, XLM-RoBERTa, GPT-4o, GPT-3.5, Few-shot Prompting, Linguistic Features

1. Introduction

1.1. Problem Overview

In the digital age, online news often uses different propaganda techniques. Propaganda, as defined by Sparkes-Vian [1], is an evolving set of methods and mechanisms that facilitate the propagation of ideas and actions. It employs rhetorical techniques to improve replication, making it a powerful tool for influencing public opinion. Propaganda is not false or immoral by its nature. Its ethical implications depend on the political, social, and technological context. Propaganda is most effective when it goes unnoticed, subtly altering readers' opinions without their awareness [2]. Therefore, detecting propaganda remains vital but also challenging to implement.

Nowadays, information spreads from many online sources. Platforms like *X* (*formerly known as Twitter*) have become vital places for sharing news and opinions. However, they have also become channels for spreading propaganda, which can influence people's thoughts and actions. As a result, it is crucial to detect propaganda, as it affects public discourse and people's decisions.

DIPROMATS 2024 organized as a part of the *Iberian Languages Evaluation Forum 2024 (IberLEF)* [3] aims to spread knowledge and research on detecting propaganda. In this study, we will present our experiments and final systems that we utilized to detect propaganda in the task organized by *DIPROMATS 2024*.

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.
†These authors contributed equally.
arkadiusz.modzelewski@pja.edu.pl (A. Modzelewski); golik.pawel@gmail.com (P. Golik); adamw@pja.edu.pl (A. Wierzbicki)
https://amodzelewski.com/ (A. Modzelewski); http://adamwierzbicki.info/ (A. Wierzbicki)
0009-0003-1169-831X (A. Modzelewski); 0009-0003-1254-6879 (P. Golik); 0000-0003-0075-7030 (A. Wierzbicki)

1.2. Task Description

DIPROMATS 2024 introduces the shared task focused on the automatic detection and characterization of propaganda techniques and narratives used by diplomats from major powers. In our experiments, we decided to focus on propaganda detection tasks. This task includes three different subtasks listed below [4]:

1. Subtask 1a: Propaganda identification
2. Subtask 1b: Propaganda characterization, coarse-grained
3. Subtask 1c: Propaganda characterization, fine-grained

Propaganda Identification Participants must develop an automatic system to determine whether a tweet contains propaganda. In this scenario, we are dealing with a binary classification task. For each instance, a short tweet text, we aim to predict one of two labels: *"false"* or *"true"* indicating the presence of propaganda [4].

Propaganda characterization, coarse-grained Systems must determine which of the four categories each tweet belongs to: *Not propagandistic*, *Appeal to commonality*, *Discrediting the opponent*, and *Loaded language*. Each tweet can be assigned to one or more categories, making it a multiclass, multilabel classification task [4].

Propaganda characterization, fine-grained Systems must classify each tweet according to the specific propaganda techniques it contains. There is one negative class and seven positive classes: *Flag Waving*, *Ad Populum/Ad Antiquitatem*, *Name Calling/Labeling*, *Undiplomatic Assertiveness/Whataboutism*, *Appeal to Fear*, *Doubt*, and *Loaded Language* [4]. This task is multiclass multilabel clasification problem.

2. Related Work

The identification of propaganda in social media and web articles has gained significant attention in recent years due to the increasing influence of online information on public opinion and political discourse [5]. Barrón-Cedeno et al. [6] proposed a model to automatically assess the level of propagandistic content on the article level. On the other hand, other approaches have introduced more fine-grained propaganda techniques detection [7, 8] and analyzed the spread of propaganda on X platform [9, 10].

Research on propaganda detection significantly intersects with persuasion detection due to the numerous shared characteristics and techniques [8]. In recent years, multiple workshops have been organized to advance the development of technologies aimed at identifying persuasion techniques [11, 12, 13, 14]. The most recent workshop (SemEval-2023 Task 3) focused on identifying 23 specific persuasion techniques in online news on paragraph level and in a multilingual setup [14]. Systems proposed during SemEval-2023 were mainly based on multilingual BERT models, such as mBERT or XLM-RoBERTa [15, 16, 17].

The detection of propaganda has also been a research focus in the most recent shared task at DIPROMATS 2023 [18]. Casavantes et al. [19] utilized BERTweet [20] and RoBERTuito [21] and aimed to improve the performance of the detection of propagandistic tweets by combining the text of tweets with contextual attributes such as their geographical origin, type of message, and emotions. UniLeon-UniBO Team utilized transfer learning between different tasks of propaganda detection [22]. Another two systems focused on employing data augmentation to improve performance in propaganda detection [23, 24]. Moreover, the best-performing system in binary propaganda classification in English was based on cascades of language models, adopting GPT-J as the backbone model [25].

Table 1
Number of tweets and authorities in Spanish and English datasets.

Region	Spanish		English	
	Tweets Count	**Authorities Count**	**Tweets Count**	**Authorities Count**
China	2,997	25	3,022	106
Russia	1,391	22	2,690	114
European Union	2,465	48	2,916	186
United States	2,738	40	3,114	216
Total	**9,591**	**135**	**12,012**	**619**

Table 2
Summary of datasets with average character and word counts. *ALL* refers to the data before our split.

Language	Dataset	Size	Positive class %	Avg. char. count	Avg. word count
English	TRAIN	7146	23.19%	255.50 ± 53.98	46.05 ± 10.98
	VALID	1262	24.88%	255.83 ± 54.31	46.22 ± 11.03
	ALL	8408	23.44%	255.55 ± 54.03	46.07 ± 10.99
Spanish	TRAIN	5202	19.30%	255.93 ± 54.26	44.40 ± 10.32
	VALID	918	20.92%	255.30 ± 53.67	44.32 ± 9.97
	ALL	6120	19.54%	255.83 ± 54.17	44.39 ± 10.29

3. Dataset

The dataset includes tweets in both Spanish and English authored by diplomats representing China, Russia, the United States, and the European Union. These tweets come from official government accounts, embassies, ambassadors, consuls, and other diplomatic profiles. The tweets were collected using the Twitter API for Academic Research and were posted between January 1, 2020, and March 11, 2021. The data contains features such as tweet ID, text, country, annotated labels, and a creation time stamp. Table 1 summarizes the presence of diplomatic authorities in the dataset [4].

The task authors split the original data into training and test sets based on time. They chose a date for each dataset that divides positive tweets into a 70/30 proportion. The 70% subset, consisting of the oldest tweets, became the training set, while the 30% subset, containing the newest tweets, became the unseen test set utilized for final systems' scores [4].

4. Our Approach

4.1. Data Preparation

Our experiments focused solely on using the tweet text and gold labels, disregarding any other columns in the dataset. In our model-building process, we included a phase for optimizing hyperparameters. To facilitate this, we divided the training data further into a new training subset and a validation subset with a ratio of 85/15. Table 2 shows the characteristics of the datasets.

4.2. Fine-tuning BERT-based Models

Our approach relied on fine-tuning pretrained BERT-based models using the labeled dataset. Fine-tuning allows the model to learn the nuances and patterns relevant to our task while retaining the general language understanding from its initial training. We employed both monolingual and multilingual pretrained models loaded from the *HuggingFace* repository:

1. **ENGLISH (ROB-EN)** - *FacebookAI/roberta-large* - the language model (355M parameters) trained on English data in a self-supervised fashion [26].

2. **SPANISH (ROB-ES)** - *PlanTL-GOB-ES/roberta-large-bne* - based on the RoBERTa large model pretrained using a large Spanish dataset, with 570GB of Spanish texts [27].
3. **BILINGUAL (XLM-BI)** - *FacebookAI/xlm-roberta-large* pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages [28].

We began with hyperparameter optimization for all the pretrained models we tested. This involved fitting N models on the training subset with gold labels and using the remaining labeled data for validation. N refers here to the number of different combinations of hyperparameter values. We selected the best model based on the F_1 score from the validation data, and this model was used for the final submission. Monolingual models were fine-tuned exclusively on tweets in a single language, while multilingual models were fine-tuned on a combination of English and Spanish tweets. Please refer to our Appendix A, which presents optimal hyperparameters of each fine-tuned model.

5. Results

5.1. Leaderboard Performance

In *DIPROMATS 2024 Task 1*, the Information Contrast Model (ICM) score determines the best propaganda categorization model, addressing the classes' hierarchical nature [29]. In presenting our results, it's important to clarify that models with the same names (e.g., **XLM-BI**) are not the same across different subtasks. For instance, the **XLM-BI** model in subtask 1a was fine-tuned specifically on subtask 1a data, while the **XLM-BI** model in subtask 1b was fine-tuned on subtask 1b data.

In subtask 1a, our best model for the English language, **ROB-EN** fine-tuned on English tweets, secured 5th place on the English leaderboard. Our bilingual **XLM-BI** model won the Spanish leaderboard and obtained 4th place on the multilingual leaderboard. In subtask 1b, the **XLM-BI** model granted us 1st position in all language categories. We also achieved first place on all language leaderboards in subtask 1c. For English, the top results were achieved by the **ROB-EN** model, while for the other leaderboards, the **XLM-BI** model prevailed. Tables 3, 4, and 5 show our final results on all subtasks.

Table 3
Our final results from the *DIPROMATS 2024 Task 1a* official leaderboard (LB).

Language	Model	ICM	F_1 score	LB rank	Winner ICM	GOLD ICM
English	ROB-EN	0.2012	0.6865	5	0.2123	0.6604
Spanish	XLM-BI	0.2187	0.7097	1	0.2187	0.6014
Bilingual	XLM-BI	0.4978	0.6896	4	0.2048	0.6323

Table 4
Our final results from the *DIPROMATS 2024 Task 1b* official leaderboard (LB).

Language	Model	ICM	F_1 macro	LB rank	Winner ICM	GOLD ICM
English	XLM-BI	0.0312	0.6219	1	0.0312	0.7014
Spanish	XLM-BI	-0.1148	0.4204	1	-0.1148	0.7535
Bilingual	XLM-BI	-0.0074	0.6029	1	-0.0074	0.6692

5.2. Results Discussion

In this workshop, our *DSHacker* team won in all categories for multiclass multilabel classification. Additionally, we secured a strong position in the binary classification subtask. Our **XLM-BI** model consistently emerged as the top solution among final submissions. However, in certain situations, monolingual models like **ROB-EN** can perform better than multilingual approaches or yield comparable results.

Table 5
Our final results from the *DIPROMATS 2024 Task 1c* official leaderboard (LB).

Language	Model	ICM	F_1 macro	LB rank	Winner ICM	GOLD ICM
English	ROB-EN	-0.0311	0.4655	1	-0.0311	0.7883
Spanish	XLM-BI	-0.0917	0.518	1	-0.0917	0.6140
Bilingual	XLM-BI	-0.0074	0.4611	1	-0.0074	0.7874

6. Few-shot Prompting with GPT Models

Another technique we explored is few-shot prompting using GPT models. In few-shot prompting, the prompt includes a brief description of the task followed by a few input-output pairs demonstrating the desired behavior. This technique allows the model to infer the patterns and rules of the task from the limited examples and generate appropriate outputs for new inputs.

We applied this approach only in the subtask 1a binary classification setting. Our experiments included OpenAI's *gpt-4o* (**GPT-4o**) and *gpt-3.5-turbo-1106* (**GPT-3.5**) generative models. We implemented the few-shot prompting technique using the OpenAI Chat Completions API. Each prediction request sent to the GPT model consisted of a list of messages presented to the model. Each message contains the role and content attribute. There are three roles available:

1. **system** message helps set the behavior of the model (assistant) by providing it context and guidelines.
2. **user** messages can provide exemplary requests for the assistant. In our case - examplary requests for a provided text's check-worthiness evaluation.
3. **assistant** messages indicate the expected output of the assistant.

Due to time constraints, we could not submit results produced by GPT models. However, we conducted post-deadline experiments and evaluated these models on the validation part of the data for binary classification from subtask 1a. In our experiments, the prompt is formatted starting with a system message that clarifies the task (See Listing 1). This is followed by alternating pairs of user and assistant messages. One pair for each few-shot example, where a user message asks whether the example's content contains propaganda, and the corresponding assistant message provides the gold label for the example, either 'Yes' or 'No' (See Listing 2). The final message following the pairs is one user message with the actual text to be classified by the model (See Listing 3). For each instance to be classified, we included four examples of few-shot prompting from the training dataset, two containing propaganda. The chosen few-shot examples were consistent in a given language. The prompt templates remained consistent for both the **GPT-4o** and **GPT-3.5** experiments.

6.1. Results on Validation Datasets

The table shows the subtask 1a F_1 scores of our models on English and Spanish validation datasets. **GPT-3.5**, using few-shot prompting, had moderate scores of 0.5193 for English and 0.5354 for Spanish. **GPT-4o** improved on these, with scores of 0.5665 for English and 0.6622 for Spanish. The multilingual model, *XLM-RoBERTa-large* (**XLM-BI**), performed better, scoring 0.7440 for English and a top score of 0.7907 for Spanish. The monolingual *RoBERTa-large* model for English (**ROB-EN**) achieved the highest score of 0.7692, while its Spanish counterpart (**ROB-ES**) scored 0.7684. Overall, fine-tuned BERT-based models outperformed GPT-based models, with multilingual and monolingual models showing similar performance.

Table 6
Results of experiments obtained on the validation dataset.

Language	Model	F_1 Score	Language	Model	F_1 Score
English	GPT-3.5	0.5193	Spanish	GPT-3.5	0.5354
	GPT-4o	0.5665		GPT-4o	0.6622
	XLM-BI	0.7440		XLM-BI	0.7907
	ROB-EN	0.7692		ROB-ES	0.7684

7. Linguistic Features for Propaganda Detection

In this section, we explore the use of StyloMetrix[1] vectors for the subtask 1a propaganda detection. StyloMetrix was sucessfully utilized in persuasion detection in Polish [17]. The study of persuasion detection significantly intersects with the study of propaganda detection due to the numerous similarities they share [30]. As a result, in our research we will explore the usage of StyloMetrix for propaganda detection in English as StyloMetrix currently does not support Spanish.

With StyloMetrix, we can create text representations that are interpretable, normalized, and reproducible [31]. By translating various aspects of linguistic features into numeric values, StyloMetrix vectors can be utilized as input for machine learning classifiers [31]. StyloMetrix quantifies many linguistic features, such as the 17 metrics created using the HurtLex lexicon. HurtLex [2] is a comprehensive lexicon encompassing offensive, aggressive, and hateful words [32]. HurtLex categorizes these words into 17 distinct groups, ranging from ethnic slurs to derogatory terms related to physical and cognitive disabilities and words associated with moral and behavioral defects [32].

In our experiments, we employ StyloMetrix vectors to predict propaganda in English tweets. The StyloMetrix vectors for English encompass a comprehensive set of 196 metrics, categorized into several groups: *Detailed grammatical forms*, *General grammar forms*, *Detailed lexical forms*, *Additional lexical items*, *Parts of speech*, *Social media*, *Syntactic forms*, *General text statistics* [31]. We utilize these text representations as features for training classical machine learning models, specifically XGBoost, LightGBM, and Logistic Regression. The models are trained on our training dataset and tested using validation data to evaluate their performance.

7.1. Results on Validation Datasets

Table 7 presents results of our experiments with classical machine learning models and linguistic features. Among the classical models, LightGBM performs the best with an F_1 score of 0.7663, followed closely by XGBoost at 0.7594. Logistic Regression trails behind significantly with a score of 0.7120, indicating that more complex models like LightGBM and XGBoost are better at capturing the nuances of the dataset. Table 6 from previus Section shows that the highest F_1 score for English validation dataset is achieved by the **ROB-EN** model, but surprisingly it is followed closely by LightGBM. Traditional machine learning models like LightGBM and XGBoost still perform robustly, showing that with well-engineered features offered by StyloMetrix vectors, they can compete closely with advanced language models in this specific task of binary propaganda detection. On the other hand, it may suggest that we should perform more comprehensive hyperparameter tunning for BERT-based models.

Table 7
Results of experiments with linguistic features and classical machine learning models obtained on the English validation dataset.

Model	F_1 Score
XGBoost	0.7594
LightGBM	0.7663
Logistic Regression	0.7120

[1]https://github.com/ZILiAT-NASK/StyloMetrix/tree/main
[2]https://github.com/valeriobasile/hurtlex

8. Conclusions

As the *DSHacker* team, we explored various techniques for propaganda detection across multiple languages and subtasks, employing both state-of-the-art pretrained BERT-based models, few-shot prompting with GPT models, and classical machine learning algorithms utilizing StyloMetrix linguistic features. Our fine-tuned BERT-based models demonstrated strong performance in the *DIPROMATS 2024 Task 1* competition. In summary, we secured 1^{st} position in 7 out of 9 categories. We won in all categories for multiclass multilabel classification and in the subtask 1a binary classification of Spanish tweets. The multilingual **XLM-BI** model consistently delivered top results, especially in multilingual and Spanish tasks. Monolingual models like **ROB-EN** also showed competitive performance, particularly for English tasks, indicating that language-specific models can sometimes outperform multilingual counterparts. Few-shot prompting with GPT models yielded moderate performance on binary propaganda classification. While **GPT-4o** beat **GPT-3.5**, both were still outperformed by fine-tuned BERT-based models. Classical machine learning models like LightGBM and XGBoost, combined with well-engineered linguistic features from StyloMetrix, performed well on the binary task of propaganda detection. LightGBM, in particular, achieved a F_1 score close to that of the best BERT-based model on the English validation dataset, highlighting the potential of classical models with rich feature sets on this specific task of propaganda detection.

References

[1] C. Sparkes-Vian, Digital propaganda: The tyranny of ignorance, Critical sociology 45 (2019) 393–409.

[2] A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, P. Nakov, Proppy: Organizing the news based on their propagandistic content, Information Processing Management 56 (2019) 1849–1864. URL: https://www.sciencedirect.com/science/article/pii/S0306457318306058. doi:https://doi.org/10.1016/j.ipm.2019.03.005.

[3] 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.

[4] 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).

[5] G. D. S. Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, P. Nakov, A survey on computational propaganda detection, arXiv preprint arXiv:2007.08024 (2020).

[6] A. Barrón-Cedeno, I. Jaradat, G. Da San Martino, P. Nakov, Proppy: Organizing the news based on their propagandistic content, Information Processing & Management 56 (2019) 1849–1864.

[7] G. Da San Martino, Y. Seunghak, A. Barrón-Cedeno, R. Petrov, P. Nakov, et al., Fine-grained analysis of propaganda in news article, 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), Association for Computational Linguistics, 2019, pp. 5636–5646.

[8] J. Piskorski, N. Stefanovitch, N. Nikolaidis, G. Da San Martino, P. Nakov, Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 3001–3022. URL: https://aclanthology.org/2023.acl-long.169. doi:10.18653/v1/2023.acl-long.169.

[9] K. Hristakieva, S. Cresci, G. Da San Martino, M. Conti, P. Nakov, The spread of propaganda by coordinated communities on social media, in: Proceedings of the 14th ACM Web Science Conference 2022, 2022, pp. 191–201.

[10] P. Vijayaraghavan, S. Vosoughi, TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3433–3448. URL: https://aclanthology.org/2022.naacl-main.251. doi:10.18653/v1/2022.naacl-main.251.

[11] G. Martino, A. Barrón-Cedeno, H. Wachsmuth, R. Petrov, P. Nakov, Semeval-2020 task 11: Detection of propaganda techniques in news articles, arXiv preprint arXiv:2009.02696 (2020).

[12] D. Dimitrov, B. B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. D. S. Martino, Semeval-2021 task 6: Detection of persuasion techniques in texts and images, arXiv preprint arXiv:2105.09284 (2021).

[13] M. Hasanain, F. Alam, H. Mubarak, S. Abdaljalil, W. Zaghouani, P. Nakov, G. Da San Martino, A. Freihat, ArAIEval shared task: Persuasion techniques and disinformation detection in Arabic text, in: H. Sawaf, S. El-Beltagy, W. Zaghouani, W. Magdy, A. Abdelali, N. Tomeh, I. Abu Farha, N. Habash, S. Khalifa, A. Keleg, H. Haddad, I. Zitouni, K. Mrini, R. Almatham (Eds.), Proceedings of ArabicNLP 2023, Association for Computational Linguistics, Singapore (Hybrid), 2023, pp. 483–493. URL: https://aclanthology.org/2023.arabicnlp-1.44. doi:10.18653/v1/2023.arabicnlp-1.44.

[14] 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), 2023, pp. 2343–2361.

[15] A. Pauli, R. Sarabia, L. Derczynski, I. Assent, Teamampa at semeval-2023 task 3: Exploring multilabel and multilingual roberta models for persuasion and framing detection, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 847–855.

[16] T. Hromadka, T. Smolen, T. Remis, B. Pecher, I. Srba, Kinitveraai at semeval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection, arXiv preprint arXiv:2304.11924 (2023).

[17] A. Modzelewski, W. Sosnowski, M. Wilczynska, A. Wierzbicki, Dshacker at semeval-2023 task 3: Genres and persuasion techniques detection with multilingual data augmentation through machine translation and text generation, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1582–1591.

[18] 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.

[19] 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).

[20] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, arXiv preprint arXiv:2005.10200 (2020).

[21] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, arXiv preprint arXiv:2111.09453 (2021).

[22] 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).

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

[24] F.-J. Rodrigo-Ginés, J. Carrillo-de Albornoz, L. Plaza, Hierarchical modeling for propaganda detection: Leveraging media bias and propaganda detection datasets (2023).

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

[26] 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). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[27] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.
[28] 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, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[29] 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.
[30] J. Piskorski, N. Stefanovitch, N. Nikolaidis, G. Da San Martino, P. Nakov, Multilingual multifaceted understanding of online news in terms of genre, framing and persuasion techniques (2023).
[31] I. Okulska, D. Stetsenko, A. Kołos, A. Karlińska, K. Głąbińska, A. Nowakowski, Stylometrix: An open-source multilingual tool for representing stylometric vectors, arXiv preprint arXiv:2309.12810 (2023).
[32] E. Bassignana, V. Basile, V. Patti, et al., Hurtlex: A multilingual lexicon of words to hurt, in: CEUR Workshop proceedings, volume 2253, CEUR-WS, 2018, pp. 1–6.

A. Optimal Hyperparameter Values

This appendix includes the optimal hyperparameter values for our best models.

Table 8
Optimal hyperparamter values used in our models. Legend: **lr** - learning_rate; **bs** - batch_size; **nte** - num_train_epochs; **ws** - warmup_steps; **wd** - weight_decay

Subtask	Model	lr	bs	nte	ws	wd
1a	ROB-EN	1×10^{-5}	16	5	200	1×10^{-2}
	ROB-ES	2×10^{-5}	16	5	200	2×10^{-2}
	XLM-BI	1×10^{-5}	16	5	200	1×10^{-3}
1b	ROB-EN	1×10^{-5}	16	5	200	2×10^{-3}
	ROB-ES	3×10^{-5}	16	5	200	1×10^{-2}
	XLM-BI	1×10^{-5}	16	5	200	2×10^{-2}
1c	ROB-EN	1×10^{-5}	16	5	200	2×10^{-2}
	ROB-ES	2×10^{-5}	16	5	200	5×10^{-2}
	XLM-BI	1×10^{-5}	16	5	200	2×10^{-2}

B. Few-shot Prompting Templates

In this appendix, we present the prompt messages included with each text classification request. For brevity, the prompts are provided only in English.

```
{
   "role": "system",
   "content": """You are an assistant who detects propaganda, manipulation
    ↪ and persuasion techniques.
   You know the definition of propaganda very well: Propaganda is the
    ↪ deliberate systematic attempt to shape perceptions and manipulate
    ↪ cognitions and direct behavior to achieve a response to further the
    ↪ desired intent of the propagandist.
   """
},
```

Listing 1: Used initial system prompt.

```
#Example 1
 # EXAMPLE 1
{
    "role": "user",
    "content": f"""Answer the question whether or not the text contains
    ↪ propaganda. Answer using only a single word Yes or No.
        TEXT: Today, I reflect on the great honor of serving the American
        ↪ people this past year, and look forward to continuing to advance
        ↪ a diplomacy true to our core values and emboldened by U.S.
        ↪ leadership that may turn our greatest challenges into our
        ↪ greatest triumphs. Happy New Year!
        """
},
{
    "role": "assistant",
    "content": "Yes"
},
# EXAMPLE 2
{
    "role": "user",
    "content": f"""Answer the question whether or not the text contains
    ↪ propaganda. Answer using only a single word Yes or No.
        TEXT: The Islamic Republic of #Iran has fundamentally failed the
        ↪ Iranian people, and I am convinced that the Iranian people know
        ↪ that. And you've seen President @realDonaldTrump make very clear
        ↪ we will continue to support the Iranian people.
        """
},
{
    "role": "assistant",
    "content": "No"
},
# EXAMPLE 3
{
    "role": "user",
    "content": f"""Answer the question whether or not the text contains
    ↪ propaganda. Answer using only a single word Yes or No.
        TEXT: The Chinese government's decision to explore its own virtual
        ↪ currency is already monumental, and if it ultimately moves
        ↪ forward it will be a global game changer. The future of global
        ↪ currencies may very well rest firmly in China's hands.
        """
},
{
    "role": "assistant",
    "content": "Yes"
},
# EXAMPLE 4
{
```

```
    "role": "user",
    "content": f"""Answer the question whether or not the text contains
    ↪  propaganda. Answer using only a single word Yes or No.
    TEXT: Over the past few years, the #US has repeatedly blocked @UN
    ↪  Security Council's statements condemning attacks on other countries'
    ↪  embassies. The US missile strike in Baghdad will only result in
    ↪  escalating tensions in the region - #Zakharova
    """
},
{
    "role": "assistant",
    "content": "No"
}
```

Listing 2: Used pairs of user and assistance prompts.

```
{
    "role": "user",
    "content": f"""Answer the question whether or not the text contains
    ↪  propaganda. Answer using only a single word Yes or No.
        TEXT: {text}
        """
}
```

Listing 3: Used final user prompt.