Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 397-407 recibido 28-06-2023 revisado 08-07-2023 aceptado 16-07-2023

Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers

Overview de DIPROMATS 2023: detección y caracterización automáticas de técnicas de propaganda en mensajes de diplomáticos y autoridades de potencias mundiales

Pablo Moral,[1, 2] **Guillermo Marco,**[1] **Julio Gonzalo,**[1]
Jorge Carrillo-de-Albornoz,[1] **Iván Gonzalo-Verdugo**[3]
[1] Universidad Nacional de Educación a Distancia
[2]Universidad Pablo de Olavide
[3]Universidad Autónoma de Madrid
{pmoral, gmarco, julio, jcalbornoz}@lsi.uned.es
ivan.gonzalo@estudiante.uam.es

Abstract: This paper presents the results of the DIPROMATS 2023 challenge, a shared task included at the Iberian Languages Evaluation Forum (IberLEF). DIPROMATS 2023 provides a dataset with 12012 annotated tweets in English and 9501 tweets in Spanish, posted by authorities of China, Russia, United States and the European Union. Three tasks are proposed for each language. The first one aims to distinguish if a tweet has propaganda techniques or not. The second task seeks to classify the tweet into four clusters of propaganda techniques, whereas the third one offers a fine-grained categorization of 15 techniques. For the three tasks we have received a total of 34 runs from 9 different teams.
Keywords: Propaganda, Digital Diplomacy, Twitter, Information Contrast Model.

Resumen: Este artículo presenta los resultados de DIPROMATS 2023, una tarea compartida incluída en el Iberian Languages Evaluation Forum (IberLEF). DIPROMATS 2023 proporciona un conjunto de datos con 12.012 tweets anotados en inglés y 9.501 tweets en español, publicados por autoridades de China, Rusia, Estados Unidos y la Unión Europea. Se proponen tres tareas para cada idioma. La primera tiene como objetivo distinguir si un tweet tiene técnicas de propaganda o no. La segunda tarea busca clasificar el tweet en cuatro grupos de técnicas de propaganda, mientras que la tercera ofrece una categorización detallada de 15 técnicas. Para las tres tareas, hemos recibido un total de 34 ejecuciones de 9 equipos diferentes.
Palabras clave: Propaganda, Diplomacia digital, Twitter, Modelo de Contraste de Información.

1 Introduction

Propaganda can be understood as "the deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist" (Jowett and O'Donnell, 2015). To this end, it involves "a set of techniques and mechanisms which facilitate the propagation of ideas and actions" (Sparkes-Vian, 2019). Propaganda's subtlety can make it a sophisticated manipulative method, as its content does not necessarily need to be false, and its features may only be identifiable through systematic long-term observation. This differentiates propaganda from disinformation, which in contrast can be exposed through objective fact-checking.

The DIPROMATS challenge has been organized for the first time with the aim of

ISSN 1135-5948 DOI 10.26342/2023-71-31

finding the best techniques to identify and categorize propagandistic tweets from governmental and diplomatic sources. Previous work that have attempted to automatically identify and classify propaganda techniques in texts have been inspired by Da San Martino et al. (2019), who classified segments within news articles depending on the technique they contained. In total, these authors initially considered 18 techniques, but in posterior related works they reduced them to 14 and even grouped them in 6 different clusters (Da San Martino et al., 2020; Da San Martino et al., 2022).

DIPROMATS 2023 is also grounded in Da San Martino et al. (2019) as we also try to detect propaganda techniques in texts, and we partially borrow their categorization. However, DIPROMATS seeks to classify tweets instead of text segments, and does not focus on news articles but on tweets published by authorities from four powers: China, Russia, United States (US) and the European Union (EU). Within the authorities considered there are accounts from government institutions and representatives, embassies, ambassadors, and other diplomatic profiles such as consuls and missions. Focusing on this type of content we intend to study governmental propaganda directly at its source.

This paper unfolds as follows. Next section describes the three tasks proposed by this challenge and their evaluation measures. Section 3 describes the dataset provided, and Section 4 an overview of the systems submitted by the participants. In Section 5 we discuss the results they obtained and finally, Section 6 ends with the main conclusions of the shared task.

2 Task

2.1 Description

DIPROMATS 2023 presented three tasks for each language, Spanish and English. Participants could choose in which task(s) and language(s) they participated:

- **Task 1: Propaganda identification**. The first task consisted of a binary classification problem. The systems had to decide whether a tweet contained propaganda techniques.
- **Task 2: Propaganda characterization, coarse**. In the second task the systems had to categorize a tweet in different groups of propaganda techniques that shared rhetorical patterns. There were four possible groups, besides a negative class: Group 0: not propagandistic, Group 1: Appeal to Commonality, Group 2: Discrediting the opponent, Group 3: Loaded Language and Group 4: Appeal to Authority.
- **Task 3: Propaganda characterization, fine-grained**. The third task asked systems to decide which of the available techniques the tweet contained. The selection of techniques was inspired by Da San Martino et al. (2019). We dismissed some of the techniques used by these authors and incorporated other techniques proposed by Johnson-Cartee and Copeland (2004) or Hobbs and McGee (2014).

For task 3 we finally considered 15 types of techniques. Two techniques belonged to **Group 1: Appeal to commonality**:

- **Ad populum / Ad antiquitatem**: the tweet appeals to the will, the tradition or the history of a community to support an argument (Weston, 2017).
  - *The leadership of the #CPC is the choice of history and of the Chinese people.*
- **Flag Waving**: the tweet includes hyperbolic praise of a nation, worships a patriotic symbol, exhibits self-praise, or portrays someone as a hero.
  - *The European Union is the best example, in the history of the world, of conflict resolution.*

Ten techniques were included in **Group 2: Discrediting the opponent**:

- **Name Calling / Labelling**: the author refers to someone or something with pejorative labels (Da San Martino et al., 2019; Institute for Propaganda Analysis, 1938).
  - *The #US is the gravest threat to global strategic security and stability.*
- **Undiplomatic Assertiveness / Whataboutism**: the tweet vilifies

an opponent, depicting their behavior as hostile, hypocritical or immoral, displaying undiplomatic contempt. This technique also includes counteraccusations to deviate the attention from sensitive issues.

  - *Just another proof that the #MediaFreedom principle is only applied to western or western-paid media. When Euro-NATO governments crack down on #Russian or Russian-language media there's zero reaction from #HumanRights apologists. Bias and double standards*

- **Scapegoating**: the tweet transfers the blame to one person, group or institution (Da San Martino et al., 2019).

  - *What has caused the current difficulties in China-UK relationship? My answer is loud and clear: China has not changed. It is the UK that has changed. The UK side should take full responsibility for the current difficulties.*

- **Propaganda Slinging**:the author accuse others of spreading propaganda, disinformation or lies (Johnson-Cartee and Copeland, 2004).

  - *Pompeo has been churning out lies wherever he goes, spreading political virus across the world.*

- **Personal attacks**: the author attacks the personal, private background of an opponent (Johnson-Cartee and Copeland, 2004).

  - Example by Johnson-Cartee and Copeland (2004): *He tries to appeal to Christian voters, but his real life is anything but Christian. He is a heavy drinker and a compulsive womanizer.*

- **Appeal to Fear**: the author either seeks to instill fear in the readers about hypothetical situations that an opponent may provoke or aims to intimidate an opponent by warning about the consequences of their actions (Johnson-Cartee and Copeland, 2004).

  - *We urge the US to stop using the Uighur Human Rights Policy Act of 2020 to harm China's interests. Otherwise, China will resolutely fight back, and the US will bear all the consequences.*

- **Absurdity Appeal**: the author characterizes the behavior of an opponent or their ideas as absurd, ridiculous or pathetic (Johnson-Cartee Copeland, 2004).

  - *Joe Biden's response to the H1N1 Swine Flu was pathetic. Joe didn't have a clue!*

- **Demonization**: the author invokes civic hatred towards an opponent, who is presented as an existential threat.

  - *Concast (@NBCNews) and Fake News @CNN are Chinese puppets who want to do business there. They use USA airwaves to help China. The Enemy of the People!*

- **Doubt**: The author casts doubt on the credibility or honesty of someone (Da San Martino et al., 2019).

  - *Growing doubts over the US government's handling of the #COVID19, e.g. When did the first infection occur in the US? Is the US government hiding something? Why they opt to blame others?*

- **Reductio ad Hitlerum**: the tweets try to persuade an audience to disapprove an action or idea from an opponent by associating it with someone or something that is hated by the audience (Teninbaum, 2009).

  - *The CPC has 90 million members, plus their families, the data has at least 270 million. Infringing these elites is directly against the Chinese people. Don't forget Hitler's evil history of persecution and massacres of German Communists and Jews.Stop NEW horrible fascists!*

Group 3: Loaded Language has only one technique called "Loaded Language", that includes hyperbolic language, evocative metaphors and words with strong emotional connotations. For example: *this monumental achievement left a tremendous mark in history!*

Finally, two more techniques are included in **Group 4: Appeal to Authority**:

- **Appeal to false authority**: the tweet includes a third person or institution to support an idea, message, or behavior for which they should not be considered as a valid expert.
  - *A voice of a Pakistani student's wife tells real situation about the coronavirus in China. Trust the Chinese Government. No panic!*
- **Bandwagoning**: The author seeks to persuade someone to join a course of action because someone else is doing it (Da San Martino et al., 2019; Hobbs and McGee, 2014).
  - *Germany took strong action today against Hizballah. We call on #EU member states to follow suit in holding Hizballah accountable.*

2.2 Evaluation measures and baselines

For the evaluation we used two metrics for classification: ICM (Amigo and Delgado, 2022) (official metric) and F1. ICM is a evaluation metric suitable for Multi-label Hierarchical classification tasks, as is the case of our tasks 2 and 3. This metric is particularly useful when the distribution of classes is highly unbalanced, both in terms of class frequency and the number of labels per item, as in the case of DIPROMATS. ICM ranges from $-$ to $+$; in order to normalize the results for a more straightforward interpretation, we rescale so that the ICM of the gold standard receives 1, and a system that always returns the least frequent class receives 0. F1 is less suited for our problem because it does not take into account the hierarchical structure of classes: a mistake between distant classes penalizes the same as a mistake between sibling classes. Also, it is insensitive to imbalanced data. However, we report it as a reference, as it is the most common metric for classification problems.

As baselines, we report a naive baseline that assigns all tweets to the most frequent class (no propaganda) and the results of a popular LLM, Roberta-base.

To obtain the baselines, we divided the training set into 90% for training and 10% for development. Then we performed a straightforward fine-tuning of the `roberta-base` model (for English) and its Spanish equivalent `PlanTL-GOB-ES/roberta-base-bne`. We modeled the problem as a multilabel classification. For the first task, we have two classes. For the second, five (the four large groups described above plus the negative class). For the third task, 13 classes, corresponding to each propaganda subtype that had examples in the training set, plus the negative class. We conducted a small grid search on the training data and picked the best hyperparameters for each task:[1]

1. Batch size: 16, 32.
2. Weight decay: 0.01, 0.1.
3. Learning rate: 1e-5, 3e-5, 5e-5.
4. Epochs: 5.

3 Dataset

DIPROMATS 2023 encompasses two annotated datasets, one composed of tweets in English and another one of tweets in Spanish. The tweets, which were collected through the Twitter API for Academic Research, were published between January 1st, 2020 and March 11th, 2021. The dataset in English contains 12012 tweets: 3022 of them were published by 106 Chinese authorities, 2960 from 114 Russian officials, 2916 from 186 authorities from the EU and 3114 tweets were posted by 216 US authorities.

The dataset in Spanish included 9591 tweets, 2997 of them published by 25 Chinese authorities, 1391 by 22 Russian authorities, 2465 tweets were published by 48 European authorities, and 40 authorities from the US provide 2738 tweets.

We split the data with a temporal criterion, choosing for each dataset the date that divides positive tweets in a 70/30 proportion. The 70%, oldest subset is the training set, and the newest 30% subset is the test set. The annotation process started by marking the fine-grained techniques detected in a tweet. If at least one technique was identified, that tweet was consequently annotated as propagandistic.

Together with the text of the tweets and their different labels, the training dataset provided information about the username that published the tweet and his/her country of origin, the tweet id, the time when the

[1]The trained baseline models and results can be found in GitHub: https://github.com/grmarco/dipromats-baselines.

tweet was posted and a sum of the retweets and likes that it received. Moreover, the type of tweet was also indicated. In total there were 10,328 organic tweets, 1,694 retweets, 1,713 quotes and 793 replies.

Four analysts contributed to the making of the annotation guide. Two of them were the annotators of both datasets, considered experts by their knowledge of international relations and philosophy of language. The other two were computer scientists that were trained for the task. The four of them initially annotated in parallel a purposive sample of 100 tweets considered propagandistic. Inter-annotator agreement was computed using Cohen's Kappa. The two main annotators had a stronger agreement (K = 0.69) when classifying into propaganda groups, whereas the agreement on the categorization of fine-grained techniques was lower (K= 0.56). This could be partially explained because of the scarcity of certain techniques in the sample. The four analysts obtained a K= 0.54 agreement for the groups and 0.39 for the fine-grained techniques.

After a detailed discussion of all tweets were there was some type of disagreement, the analysts revised the annotation criteria. The two main annotators then manually classified in parallel another representative sample of 231 tweets. In this second annotation they had a agreement of K= 0.86 when detecting if a tweet contained any technique and K= 0.81 when identifying the first three propaganda groups (the fourth group only had one example in the sample). Moreover, the agreement was K= 0.8 when annotating the six techniques that were represented more than once in the sample.

The manual annotation revealed that one of the challenging aspects of this task is that the training dataset in both languages is very unbalanced. First, the binary classification, corresponding to task 1, shows a considerably higher proportion of non-propagandistic tweets: only 23.4% of the tweets in English and 19.6% of tweets in Spanish contained at least one technique. At the group level, even if groups 1, 2 and 3 appeared frequently, only 4 tweets in English and 6 in Spanish had techniques associated with Group 4, Appeal to Authority (see Figure 1). In Spanish, group 2, Discrediting the Opponent had almost as much frequency as the other three groups combined.

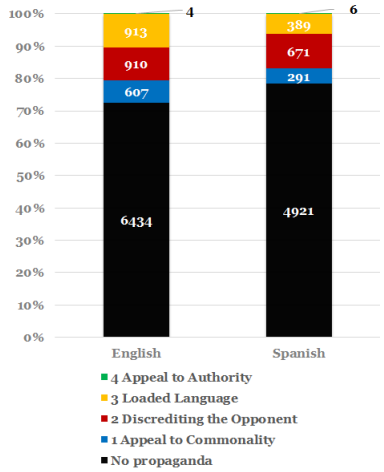

Figure 1: Number of tweets that contained each of the propaganda groups in Spanish and English.

At a fine-grained level, the techniques within the groups were also unevenly distributed as Figure 2 and Figure 3 show. In both languages, Flag Waving was much more used than Ad Populum / Ad Antiquitatem in group 1. In group 2 there were two techniques that did not have any example in the training dataset: Reductio ad Hitlerum and Personal Attacks. Other techniques, such as Scapegoating, barely appeared. On the contrary, Undiplomatic Assertiveness / Whataboutism had a very high occurrence in both languages.

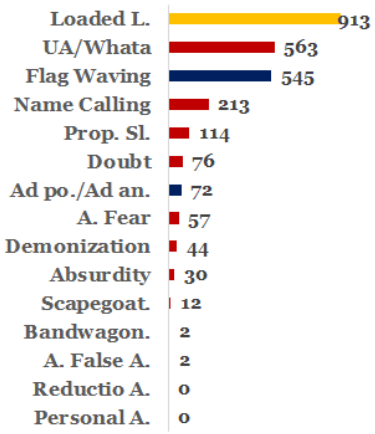

Figure 2: Number of tweets in English containing each technique. Color indicates belonging to different groups.

4 *Overview of the systems*

In total, 28 groups registered for the task, out of which 9 teams from 4 countries submitted a total of 34 runs (each participant was al-

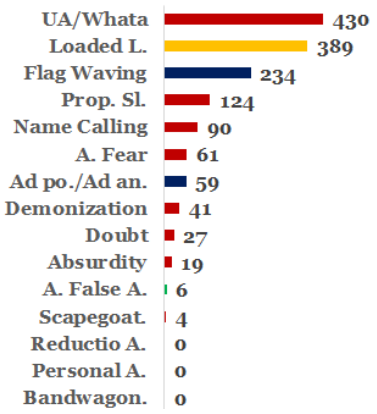

Figure 3: Number of tweets in Spanish containing each technique. Color indicates belonging to different groups.

lowed to submit up to five runs, where each run might contain results for part of all of the tasks in one or both languages). For task 1, 18 runs were submitted in Spanish by 6 different teams, whereas 9 teams submitted 30 runs in English. For task 2 and 3 we received 17 runs in Spanish from 5 different teams and 28 runs in English from 7 participants. 7 out of the 9 teams that sent runs also submitted working notes describing their systems.

Concerning the classification approaches, all seven teams employed some sort of transformer architecture for at least one of the tasks. Six teams relied on BERT-based (Devlin et al., 2018) approaches, being RoBERTa (Liu et al., 2019) the most popular one. Alternatively, one team, *Mario*, used the open-source GPT-J model. Traditional machine learning methods such as Nearest Neighbors (kNN) and Boolean bag-of-words were also applied in task 2 by team *PropaLTL*. Next, we briefly describe the approaches presented by each team.

ELiRF-VRAIN participated in the three tasks for both languages. This team used data augmentation to increase the number of samples by translating Spanish samples into English and vice versa. They employed BETO (Cañete et al., 2020) for tweets in Spanish and an updated, RoBERTa-based version of TimeLM (Loureiro et al., 2022) for tweets in English. For the fine-tuning process they created a Discrepancy Correction Procedure to prevent labeling inconsistencies.

Mario just participated in task 1 and only in English. Their approach was based on a system of cascades of language models, adopting GPT-J as the backbone model for all the experiments.

NL4IA participated in the three tasks, but only in English. They employed data augmentation by adding textual information about the sentiment, the interactions and the author's country of the tweet. Then, they applied RobertaforSequenceClassification (Wolf et al., 2020) as the pre-trained model for the classification task.

PropaLTL took part in the three tasks and in both languages, although they focused primarily on the binary propaganda identification task. Their approach consisted in adding contextual information to the tweet text, considering its sentiment, the tweet type and the country of the author. For task 1 they used BERTweet (Nguyen, Vu, and Tuan Nguyen, 2020) and RoBERTuito (Pérez et al., 2022), whereas for Task 2, as mentioned above, they decided to exploit a k-Nearest Neighbors (kNN) classifier, together with a Boolean bag-of-words representation.

UMUteam also participated in the three tasks for both languages. However, their focus was mainly on the third task, the fine-grained classification, inferring thus the labels for tasks 2 and 1. Their methodology involved a typical machine-learning pipeline that consisted in cleaning the dataset, extracting features from the documents and train and evaluate a variety or BERT-based models.

UnedMediaBiasTeam participated in the three tasks and in both languages. This team developed a three-stage hierarchical model using a fine-tuned XLM-RoBERTa (Conneau et al., 2020) and leveraging the provided data as well as data from the SemEval'23 task 3 dataset (Piskorski, J. et al.,), and from the MBIC (Spinde et al., 2021a) and BABE (Spinde et al., 2021b) datasets.

Finally, *UniLeon-UniBO* followed a bottom-up strategy to address the three tasks in both languages. On the basis of the decision in task 3, the systems determined the label in task 2 and 1. Their models were built on top of RoBERTa, and in the fine-tuning process they resorted to the Propaganda Techniques Corpus (Da San Martino et al., 2020) to increase the number of instances for certain techniques. They selected and extracted the sentences from this corpus that contained the same (or similar) techniques to those in

DIPROMATS. They also made the model aware of the author's country by attaching it as a contextual feature in the text of the tweet.

5 System results

The three tasks were evaluated independently. Teams were ranked by the ICM result they obtained. A detailed classification that includes all the runs submitted by participants and the F1 of the two classes of task 1 is available at the DIPROMATS website.[2]

5.1 Task 1

18 different runs were submitted for task 1 in Spanish and 30 in English. 16 runs were considered for the bilingual evaluation. As Table 2 shows, *PropaLTL* achieved the best score in Spanish. Their best run, the third one they submitted, incorporated as a contextual feature the type of the tweet. In English, *Mario* obtained the best result in the only run this team submitted, following a system of cascade based on GPT-J. The bilingual ranking was also topped by *PropaLTL*, whose five runs achieved the best scores. Run 4, incorporating information about the emotion of the text, was the top performer.

In task 1 the differences in the performance of the top systems in different languages are not significant. The best run in Spanish is two points higher in terms of ICM, and their F1 is practically identical.

5.2 Task 2

For task 2 there were 17 runs in Spanish, 28 in English and 15 for both languages. The best score in Spanish was achieved by *UniLeon-UniBO* using the RoBERTa architecture from the BERTIN project (De la Rosa et al., 2022). They started conducting a Twitter-oriented preprocessing keeping emoticons, emojis, and other features, while splitting hashtags and user mentions into words. They also incorporated information about the country of origin of the author. Finally, they enriched the traning dataset with a dataset that considered similar propaganda techniques to identify propagandistic segments in news articles.

With this approach, *UniLeon-UniBo* ranked second in English. In this language, top-performer *NL4IA*, which relied on RoBERTa-Large, also enriched the tweet text with contextual information: the sentiment and interaction metrics of the tweet. In the bilingual evaluation, *UMUteam* achieved the best result based on a ensemble learning with the mode of the predictions.

In this task, the difference in performance between Spanish and English is larger. The best run in English obtained slightly better results than the best Spanish run in terms of ICM, but in terms of F1 English results are 12 points higher.

5.3 Task 3

17 runs in Spanish and 28 in English participated in task 3. 15 runs were considered for the bilingual classification. As in task 2, *UniLeon-UniBo* attained the best score in Spanish. In English, the highest score was achieved again by *NL4IA*, with a run that attached information on the country of origin of the authorities to the text, a strategy that was also followed by the top-performer in Spanish *UniLeon-UniBo*.

The ICM of the best systems in task 3 remained close between Spanish and English. However, the difference between the performances in both languages kept widening in terms of F1 score. A 20 point-gap separates the best result in English from the best one in Spanish (see Table 2 and Table 3).

Note that it is not trivial to perform better than a straightforward fine-tuning of a standard LLM: The Roberta-base baseline is beaten typically by two or three systems, depending on the task.

6 Conclusions

This paper presented the results of the first edition of DIPROMATS, which challenged participants to automatically detect and classify propagandistic messages from public representatives. This shared task provided an original annotated dataset and a novel categorization, proposing a framework that allowed to test systems in two different languages and in different levels of granularity.

The approaches adopted by participants were very diverse. Generally, the best systems incorporated some kind of data augmentation that included contextual information in the message analyzed. Some successful approaches conducted bottom-up strategies that focused on the fine-grained level to resolve the more coarse-grained tasks. Conversely, the team *PropaLTL*, that decided to

[2] https://sites.google.com/view/dipromats2023/results.

Both languages								
Task 1			**Task 2**			**Task 3**		
Team	**ICM**	**F1**	**Team**	**ICM**	**F1**	**Team**	**ICM**	**F1**
Gold	1	1	Gold	1	1	Gold	1	1
PropaLTL	**0.8196**	**0.7953**	**UMUteam**	**0.9146**	0.4815	**ELiRF-VRAIN**	**0.9122**	**0.3616**
UnedMBT	0.8048	0.777	ELiRF-VRAIN	0.9139	**0.4838**	UMUteam	0.9115	0.3284
UMUteam	0.8041	0.7734	UnedMBT	0.9129	0.4639	UnedMBT	0.9082	0.2793
ELiRF-VRAIN	0.8004	0.7732	PropaLTL	0.9008	0.3824	PropaLTL	0.871	0.0674
INGEOTEC	0.7639	0.7365						
Baseline max freq.	0.6647	0.4565	Baseline max freq.	0.8665	0.1826	Baseline max-freq.	0.8704	0.0652
Baseline roberta-base.	0.8256	0.8024	Baseline roberta-base.	0.9229	0.5223	Baseline roberta-base.	0.9184	0.3225

Table 1: Results of tasks 1, 2 and 3 for tweets in Spanish and English (best run).

Spanish								
Task 1			**Task 2**			**Task 3**		
Team	**ICM**	**F1**	**Team**	**ICM**	**F1**	**Team**	**ICM**	**F1**
Gold	1	1	Gold	1	1	Gold	1	1
PropaLTL	**0.8421**	**0.8089**	**UniLeon**	**0.9123**	0.4301	**UniLeon**	**0.9043**	0.2788
UMUteam	0.8275	0.7887	UMUteam	0.9118	0.4164	ELiRF-VRAIN	0.9035	**0.3628**
UniLeon	0.825	0.7864	ELiRF-VRAIN	0.9098	**0.4578**	UMUteam	0.9017	0.3414
ELiRF-VRAIN	0.8203	0.7815	UnedMBT	0.9054	0.4079	UnedMBT	0.8946	0.2733
UnedMBT	0.8176	0.7757	PropaLTL	0.8892	0.3761	PropaLTL	0.8622	0.0789
INGEOTEC	0.792	0.7485						
Baseline max freq.	0.6852	0.4531	Baseline max freq.	0.8586	0.1812	Baseline max-freq.	0.8598	0.0697
Baseline roberta-base.	0.8448	0.8121	Baseline roberta-base.	0.9174	0.4707	Baseline roberta-base.	0.9053	0.3105

Table 2: Results of tasks 1, 2 and 3 for tweets in Spanish (best run).

focus mostly on the binary classification (task 1), also obtained remarkable results in that task. Also, it is remarkable that the baseline, a roberta-base carefully trained, obtains results that are acceptable from the outset and not trivial to surpass.

As expected, systems achieved worse performances as the complexity of the task increased. This partially explains the meaningful differences among the results of the three tasks. The degree of difficulty also seem to have an impact in the performance of the systems when dealing with different languages: the more complex the task, the wider the gap between English and Spanish models.

The use of ICM, a novel measure which is theoretically sound in terms of multiclass, multilabel hierarchical classification, has been a mixed experience. It is obviously better suited for our tasks 2 and 3, and therefore better reflects the true behaviour of systems. But, at least with the normalization schema that we have applied, it seems to be less discriminative than F1.

Results confirm that automated propaganda detection is a challenging exercise that

English								
Task 1			**Task 2**			**Task 3**		
Team	**ICM**	**F1**	**Team**	**ICM**	**F1**	**Team**	**ICM**	**F1**
Gold	1	1	Gold	1	1	Gold	1	1
Mario	**0.8202**	**0.809**	**NL4IA**	**0.9392**	**0.5591**	**NL4IA**	**0.9247**	**0.4838**
PropaLTL	0.818	0.8062	UniLeon	0.9356	0.549	UniLeon	0.9205	0.4405
UniLeon	0.8132	0.8011	UnedMBT	0.926	0.4879	ELiRF-VRAIN	0.9085	0.3768
NL4IA	0.808	0.7953	ELiRF-VRAIN	0.9256	0.5058	UnedMBT	0.9073	0.3229
UnedMBT	0.7924	0.7774	UMUteam	0.925	0.4976	UMUteam	0.9064	0.3253
UMUteam	0.7853	0.7677	PropaLTL	0.9148	0.3866	PropaLTL	0.8628	0.0721
ELiRF-VRAIN	0.7848	0.7709	IIA-CSIC	0.5971	0.1689	IIAA-CSIC	0.6187	0.067
INGEOTEC	0.736	0.7255						
IIA-CSIC	0.7023	0.6981						
Baseline max freq.	0.6456	0.4600	Baseline max freq.	0.8870	0.1840	Baseline max-freq.	0.8636	0.0767
Baseline roberta-base.	0.8066	0.7938	Baseline roberta-base.	0.9327	0.5339	Baseline roberta-base.	0.9176	0.3418

Table 3: Results of tasks 1, 2 and 3 for tweets in English (best run).

has still room for improvement, particularly when distinguishing among different techniques. Future work must also address the unbalanced distribution of categories, which may have constituted an obstacle for the training and testing processes. All in all, the enriching diversity of approaches submitted for the first edition of DIPROMATS has contributed to the literature on automated propaganda detection by providing valuable indications on where to direct ensuing efforts.

Acknowledgements

This work was partially supported by the Spanish Ministry of Science and Innovation under the projects "FairTransNLP: Midiendo y Cuantificando el sesgo y la justicia en sistemas de PLN" (PID2021-124361OB-C32), and "Desinformación y agresividad en Social Media: bias, controversia y veracidad" (PGC2018-096212-B-C32). This work has also been partially financed by the European Union (NextGenerationEU funds) through the "Plan de Recuperación, Transformación y Resiliencia", by the Ministry of Economic Affairs and Digital Transformation and by UNED. However, the points of view and opinions expressed in this document are solely those of the authors and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be considered responsible for them.

Guillermo Marco is supported by the Spanish Ministry of Science and Innovation under the grant FPU20/07321 and he is also a postgraduate fellow of the City Council of Madrid at the Residencia de Estudiantes (2022–2023).

References

Amigo, E. and A. Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland, May. Association for Computational Linguistics.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020(2020):1–10.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale, April. arXiv:1911.02116 [cs].

Da San Martino, G., A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Da San Martino, G., P. Nakov, J. Pirskoski, and N. Stefanovitch. 2022. SemEval2023 shared task on "Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup".

Da San Martino, G., S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov. 2019. 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)*, pages 5636–5646, Hong Kong, China, November. Association for Computational Linguistics.

De la Rosa, J., E. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, and M. Grandury. 2022. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural*, pages 13–23.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Publisher: arXiv Version Number: 2.

Hobbs, R. and S. McGee. 2014. Teaching about Propaganda: An Examination of the Historical Roots of Media Literacy. *Journal of Media Literacy Education*, 6(2).

Institute for Propaganda Analysis. 1938. *Propaganda Analysis: Volume I of the Publications of the Institute for Propaganda Analysis*. Institute for Propaganda Analysis, Inc., New York, 1 edition.

Johnson-Cartee, K. S. and G. Copeland. 2004. *Strategic political communication: rethinking social influence, persuasion, and propaganda*. Communication, media, and politics. Rowman & Littlefield, Lanham, Md.

Jowett, G. and V. O'Donnell. 2015. *Propaganda & persuasion*. SAGE, Thousand Oaks, Calif, sixth edition edition.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July. arXiv:1907.11692 [cs].

Loureiro, D., F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Nguyen, D. Q., T. Vu, and A. Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Piskorski, J., Stefanovitch, N., Da San Martino, G., and Nakov, P. 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*, Toronto, Canada.

Pérez, J. M., D. A. Furman, L. Alonso Alemany, and F. M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France, June. European Language Resources Association.

Sparkes-Vian, C. 2019. Digital Propaganda: The Tyranny of Ignorance. *Critical Sociology*, 45(3):393–409, May.

Spinde, T., L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay. 2021a. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics, May. arXiv:2105.11910 [cs].

Spinde, T., M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa. 2021b. Neu-

ral Media Bias Detection Using Distant Supervision With BABE – Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177. arXiv:2209.14557 [cs].

Teninbaum, G. H. 2009. Reductio ad Hitlerum: Trumping the judicial Nazi card. *Mich. St. L. Rev.*, page 541. Publisher: HeinOnline.

Weston, A. 2017. *A rulebook for arguments*. Hackett Publishing Company, Inc, Indianapolis ; Cambridge, fifth edition edition.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.