Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 347-358 recibido 09-07-2024 revisado 17-07-2024 aceptado 30-07-2024

Overview of DIPROMATS 2024: Detection, Characterization and Tracking of Propaganda in Messages from Diplomats and Authorities of World Powers

Overview de DIPROMATS 2024: detección, caracterización y seguimiento de la propaganda en mensajes de diplomáticos y autoridades de potencias mundiales

Pablo Moral, Jesús M. Fraile, Guillermo Marco, Anselmo Peñas, Julio Gonzalo
Universidad Nacional de Educación a Distancia
{pmoral, jfraile, gmarco, anselmo, julio}@lsi.uned.es

Abstract: This paper summarizes the findings of DIPROMATS 2024, a challenge included at the Iberian Languages Evaluation Forum (IberLEF). This second edition introduces a refined typology of techniques and a more balanced dataset for propaganda detection, alongside a new task focused on identifying strategic narratives. The dataset for the first task includes 12,012 annotated tweets in English and 9,501 in Spanish, posted by authorities from China, Russia, the United States, and the European Union. Participants tackled three subtasks in each language: binary classification to detect propagandistic tweets, clustering tweets into three propaganda categories, and fine-grained categorization using seven techniques. The second task presents a multi-class, multi-label classification challenge where systems identify which predefined narratives (associated with each international actor) tweets belong to. This task is supported by narrative descriptions and example tweets in English and Spanish, using few-shot learning techniques. 40 runs from nine different teams were evaluated.
Keywords: Propaganda, Narratives, Digital Diplomacy, Information Contrast Model.

Resumen: Este trabajo presenta DIPROMATS 2024, una tarea compartida incluída en IberLEF. Esta segunda edición introduce una tipología refinada de técnicas y un conjunto de datos más equilibrado para la detección de propaganda, además de una nueva tarea para la detección de narrativas estratégicas. El dataset de la primera tarea incluye 12.012 tweets en inglés y 9.501 en español de autoridades de China, Rusia, Estados Unidos y la Unión Europea. Los participantes abordaron tres subtareas por idioma: clasificación binaria de tweets propagandísticos, su agrupación en tres categorías de propaganda y su categorización en siete técnicas. La segunda tarea consiste en una clasificación multi-clase y multi-etiqueta para identificar a cuáles de las narrativas predefinidas pertenecen los tweets, siguiendo descripciones y ejemplos en inglés y español (aprendizaje few-shot). En total fueron evaluadas 40 ejecuciones de nueve equipos diferentes.
Palabras clave: Propaganda, Narrativas, Diplomacia Digital, Modelo de Contraste de Información.

1 Introduction

Building upon the foundation set by DIPROMATS 2023 (Moral et al., 2023), the inaugural edition of this shared task, DIPROMATS 2024 has been launched to refine the challenge and explore new dimensions of propaganda analysis. This second edition not only introduces adjustments to the propaganda detection and characterization task proposed last year but also includes a new task focused

ISSN 1135-5948 DOI 10.26342/2024-73-26

on narratives detection. 28 teams registered and 9 submitted results for this year's challenge.

Propaganda involves a deliberate, systematic and politically motivated attempt to shape perceptions (Jowett and O'Donnell, 2015). It "deliberately misrepresent symbols, appealing to emotions and prejudices and bypassing rational thought, to achieve a specific goal of its creators" (Bolsover and Howard, 2017) employing techniques and mechanisms that facilitate its propagation (Sparkes-Vian, 2019; Da San Martino et al., 2019). Inspired by previous works such as Da San Martino et al. (2019), the first edition of DIPROMATS proposed a typology of 15 propaganda techniques grouped in 4 different clusters according to their rhetorical properties. DIPROMATS 2024 introduces a revised version of this taxonomy to address two significant issues from the previous edition. Firstly, it aims to balance the dataset, which was uneven due to manual annotation based on the original typology. Secondly, it addresses the similarity of certain techniques within groups, which negatively impacted system performance.

In addition, DIPROMATS 2024 proposes to extend the challenge to obtain a more comprehensive diagnose of propagandistic endeavors. Besides looking into the manipulative techniques and the role of rhetoric in propaganda, in this edition we aim at the big picture: the narratives that underlie the persuasion efforts. Propaganda studies show that political beliefs are constructed from carefully built narratives which weave together plausible suggestions, half-truths and manipulative assertions (Richards, 2023). Narratives are widely considered a persuasive form of communication as they provide an appealing causal sequence to rationalize events. People make sense of what is happening by absorbing experiences into episodic narratives; they retain in their individual memory a continuous, logical sequence made of discontinuous events (Bolt, 2012). On X (formerly Twitter), tweets function as concise narrative snippets that progressively form an interdependent a meaningful storyline (Moral, 2024).

Narratives are causally connected sequences of events that are selected and evaluated as meaningful for a particular audience (Riessman, 2008). In international relations, international actors create strategic narratives to "construct a shared meaning of the past, present, and future of international politics to shape the behavior of domestic and international actors" (Miskimmon, O'Loughlin, and Roselle, 2013). Some authors consider strategic narratives to be a powerful form of propaganda, as they obscure their persuasive intent by embedding messages in a narrative drama (Colley, 2020).

Previous work on narrative detection and analysis commonly relied on clustering methods that infer narratives from words co-occurrence. However, we believe that this approach has inherent limitations. Narratives do not necessarily share lexical features as what defines a narrative is not the lexical similarity but the presence of common narrative elements: agents, agencies, scenes, act and, notably, purpose.

Thus, the primary goal of this new task is to explore how narratives can be used to group messages that share a communicative intention even if their lexical appearance is very different. To the best of our knowledge, ours is the first work that links the automated detection of narratives with political propaganda and, in a wider sense, with information campaigns. It is also the first one that tries to automatedly detect narratives in tweets of public representatives.

Therefore, DIPROMATS 2024 proposes a two-fold challenge that pursues (1) the automated identification and categorization of propaganda techniques in tweets and (2) the automated identification of narratives in individual tweets. These tasks are independent, allowing participants to choose whether to participate in one or both. The tasks involve analyzing tweets from authorities in Russia, China, the US, and the EU, but each task utilizes different datasets, evaluation measures, and methodologies. Consequently, this paper is divided into two main sections. Section 2 provides an overview of Task 1, while Section 3 details Task 2.

2 Task 1: Automatic Detection and Catergorization of Propaganda Techniques

2.1 Description

DIPROMATS 2024 comprised three subtasks for each language, Spanish and English. Participants could choose in which subtask and

language they participated.

- **Subtask 1A: Propaganda identification**. The systems had to decide whether a tweet contained propaganda techniques in a binary classification problem.
- **Subtask 1B: Propaganda characterization, coarse**. Systems were required to categorize a tweet in three different groups of propaganda techniques and a negative class: *Group 0: not propagandistic*, *Group 1: Appeal to Commonality*, *Group 2: Discrediting the opponent*, and *Group 3: Loaded Language*.
- **Subtask 1C: Propaganda characterization, fine-grained**. Tweets had to be classified according to the propaganda techniques they contained. There are a negative class and seven possitive classes. *Flag Waving, Ad Populum / Ad antiquitatem*, *Name Calling/Labelling*, *Undiplomatic Assertiveness / Whataboutism*, *Appeal to Fear*, *Doubt*, and *Loaded Language*.

These three subtasks correspond to the three available tasks in DIPROMATS 2023. The typology of techniques[1] was synthesized to achieve a more balanced distribution of techniques. Last year's *Group 4: Appeal to Authority* was removed due to its minimal presence in the dataset (only 10 instances in the training dataset). Within *Group 2: Discrediting the opponent*, the number of techniques has been reduced from 10 to 4. This reduction involved eliminating two techniques, *Reductio ad Hitlerum* and *Personal Attacks*, as there were no occurrences in the training dataset. Furthermore, to consolidate techniques that share similar characteristics and often co-occur in tweets, several techniques (*Propaganda Slinging*, *Absurdity Appeal*, *Scapegoating*, and *Demonization*) have been combined under (*Undiplomatic Assertiveness / Whataboutism*.

2.2 Dataset

The revised distribution of techniques resulted in an updated version of the annotated datasets in both English and Spanish. The data source remains the same as in the previous edition. The tweets included, collected via the Twitter API for Academic Research, were published between January 1, 2020, and March 11, 2021, by government institutions, representatives, embassies, ambassadors, and other diplomatic profiles. The English dataset comprises 12,012 tweets, with 3,022 posted by 106 Chinese authorities, 2,960 by 114 Russian diplomats, 2,916 by 186 EU officials, and 3,114 by 216 US authorities. In Spanish, 2,997 tweets were posted by 25 Chinese authorities, 1,391 by 22 Russian authorities, 2,465 by 48 EU authorities, and 2,738 by 40 US authorities, totaling 9,591 tweets in this language.

The data was split based on a temporal criterion, selecting for each dataset the date that divides positive tweets in a 70/30 proportion. The oldest 70% forms the training set, while the newest 30% forms the test set. Along with the tweet text and their various labels, the training dataset includes information about the username that published the tweet, their country of origin, the tweet ID, the posting time, and a combination of the retweets and likes it received. Additionally, the type of tweet is indicated. In total, there were 10,328 organic tweets, 1,694 retweets, 1,713 quotes, and 793 replies. The annotation process and inter-annotator agreement remain as detailed in Moral et al. (2023).

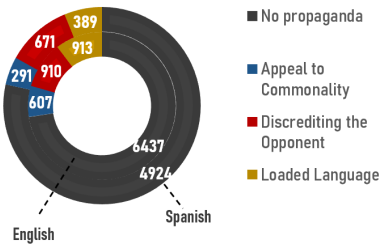

Figure 1: Number of tweets containing each group of techniques

In the training set, only 23.4% of the tweets in English and 19.6% of the tweets in Spanish contained propaganda techniques. The frequency of the three groups of techniques varied between English and Spanish, with *Group 1: Appeal to Commonality* being the least frequent. Regarding the distribution of techniques, removing the least frequent ones led to a more balanced dataset. Despite notable disparities between the most and least used techniques at a detailed level, each type had at least dozens of instances in the training dataset, unlike the technique dis-

[1]Details are fully provided at https://sites.google.com/view/dipromats2024/task-1/typology-of-techniques and in Moral et al. (2023).

tribution in DIPROMATS 2023.

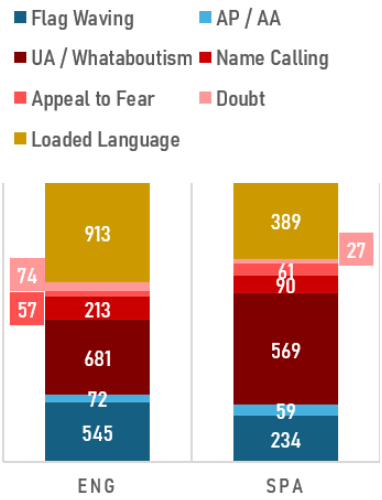

Figure 2: Number of tweets containing each type of techniques

2.3 Evaluation measures and baselines

For the evaluation, we utilized two metrics for classification: ICM (Amigo and Delgado, 2022) (the official metric) and F1. ICM is tailored for multi-label hierarchical classification tasks, which apply to our tasks 2 and 3. This metric is advantageous when dealing with highly imbalanced class distributions, both in terms of class frequency and the number of labels per item, as observed in DIPROMATS. ICM scores range from $-\infty$ to $+\infty$; for easier interpretation, we normalize the results so that the gold standard is 1 following the formula (ICM score + gold standard)/(2*gold standard). Traditional F1 score is less suitable for our problem because it does not consider the hierarchical nature of the classes: it penalizes errors between distant classes as heavily as errors between sibling classes, and it is not sensitive to imbalanced data. Nevertheless, we include F1 as a reference point, given its widespread use in classification problems.

We provide two baselines: a naive baseline that assigns all tweets to the most frequent class (no propaganda) and best performing baseline of the ODESIA leaderboard[2] for DIPROMATS 2023: `XLM-RoBERTa-Large`. To establish the baselines, we divided the training set into 90% for training and 10% for development. We then fine-tuned the `XLM-RoBERTa-Large` model for English and Spanish. A small grid search on the training data helped us determine the best hyperparameters. We tested batch sizes of 16 and 32, weight decays of 0.01 and 0.1, learning rates of 1e-5, 3e-5, and 5e-5, and a fixed 5 epochs.

2.4 Systems overview and results

Eight teams from six different countries participated in Task 1, with seven teams successfully submitting working notes explaining their approaches. Participants could submit up to five runs and each run could include results in Spanish, English or both languages. In total, 35 runs were submitted for the three subtasks.

The participants employed a wide range of approaches. Team *DSHacker* focused solely on tweet text, using fine-tuned monolingual and multilingual BERT-based models (Devlin et al., 2018). *IIIA* utilized a pretrained TwHIN-BERT encoder (Zhang et al., 2023) and incorporated sociopolitical contextual data, including the 7-day rolling average of armed conflicts and COVID-19-related deaths. *PropaLTL* used data augmentation, adding emotion information to tweets. They employed BERT-based classifiers such as BERTweet (Nguyen, Vu, and Tuan Nguyen, 2020) for English and RoBERTuito (Pérez et al., 2022) for Spanish, translating tweets to enable cross-lingual experiments. *UC3M-LCPM* used various BERT-like, GPT-like, and XLNet-like models, applying data augmentation to balance the training dataset by paraphrasing tweets, and used a model pre-trained on the Google PAWS dataset (Alisetti, 2024) for English sentences. For Spanish sentences, they translated them to English for paraphrasing and then back to Spanish. *UMUTeam* extracted linguistic features using UMUTextStats (García-Díaz et al., 2022) from different BERT-based models to train ten models for each language. *VerbaNex-AI* combined feature extraction from TF-IDF and transformers, employing regularization techniques such as class balancing and k-fold cross-validation. *Victor Vectors* tested RoBERTa (Liu et al., 2019), BETO (Cañete et al., 2020), and multilingual BERT, focusing on dataset imbalances and using data augmentation strategies like translating and paraphrasing text samples between Spanish and English.

The three subtasks of Task 1 were evalu-

[2]https://leaderboard.odesia.uned.es/

English								
Subtask 1A			**Subtask 1B**			**Subtask 1C**		
Team	**ICM**	**F1**	**Team**	**ICM**	**F1**	**Team**	**ICM**	**F1**
Gold	1	1	Gold	1	1	Gold	1	1
Victor Vectors	**0.6607**	**0.813**	**DSHacker**	**0.5222**	0.6219	**DSHacker**	**0.4803**	**0.4655**
UMUTeam	0.6577	0.8117	Victor Vectors	0.5066	**0.6288**	Victor Vectors	0.4739	0.4488
DSHacker	0.6523	0.8074	UMUTeam	0.4759	0.5877	UMUTeam	0.4425	0.408
PropaLTL	0.6364	0.7994	UC3M-LCPM	0.3293	0.4628			
UC3M-LCPM	0.5957	0.7759						
IIIA	0.5666	0.7568						
Baseline max freq.	0.3048	0.4532	Baseline max freq.	0.2174	0.2266	Baseline max-freq.	0.1986	0.1133
Baseline xlm-roberta-large	0.6252	0.7935	Baseline xlm-roberta-large	0.5530	0.6555	Baseline xlm-roberta-large	0.5303	0.5344

Table 1: Results of subtasks 1A, 1B and 1C for tweets in English (best run).

ated separately, with teams ranked based on their ICM results. A detailed classification of all participant submissions and their F1 scores is available on the DIPROMATS website.[3] Tables 1, 2 and 3 display the rankings for each subtask based on the best run submitted by each team.

Subtask 1A

For Subtask 1A, 26 runs were submitted in English, 30 in Spanish, and 33 for the bilingual evaluation. *Victor Vectors* achieved the highest scores in both ICM and F1 metrics in English and the bilingual evaluation, using RoBERTa-Large on an augmented corpus that included paraphrased examples from the *Ad Populum / Ad Antiquitatem* and *Appeal to Fear* classes in Subtask 1C. These predictions were used to infer labels for the other subtasks. In Spanish, *DSHacker* achieved the best performance using FacebookAI/xlm-roberta-large pre-trained on 2.5TB of filtered Common-Crawl data containing 100 languages (Conneau et al., 2020). Compared to the previous edition, top performers in this subtask improved by 1 F1 point in the bilingual evaluation, more than 2 points in Spanish, and 0.14 points in English. Notably, the best English run had a lower F1 score than the best runs in Spanish and the bilingual evaluations.

Subtask 1B

For Subtask 1B, 15 bilingual runs and 14 runs in English and Spanish were evaluated. *DSHacker* was the top performer in all three evaluations, using the same approach as in Subtask 1A. Results significantly improved from the previous edition, with the best English run improving by more than 6 points, the best Spanish run by 8 points, and the bilingual evaluation by 12 points. Unlike Subtask 1A, the best English run outperformed the best Spanish run by more than 10 points.

Subtask 1C

DSHacker also excelled in Subtask 1C, with the highest scores among the 10 bilingual runs and the 9 runs in English and Spanish. While the Spanish and bilingual approaches were consistent with the previous subtasks, the best English run used FacebookAI/roberta-large with 355M parameters trained on English data in a self-supervised manner. The best English run was slightly over 1 point below the previous year's top performer. However, the Spanish and bilingual evaluations improved by over 12 and 10 points, respectively. The gap between English and Spanish in this subtask

[3] https://sites.google.com/view/dipromats2024/task-1/results.

Spanish								
Subtask 1A			**Subtask 1B**			**Subtask 1C**		
Team	**ICM**	**F1**	**Team**	**ICM**	**F1**	**Team**	**ICM**	**F1**
Gold	1	1	Gold	1	1	Gold	1	1
DSHacker	**0.6818**	**0.8306**	**DSHacker**	**0.4990**	**0.518**	**DSHacker**	**0.4238**	**0.4204**
Victor Vectors	0.6625	0.8207	Victor Vectors	0.3981	0.4643	Victor Vectors	0.3566	0.3936
PropaLTL	0.6241	0.798	UMUTeam	0.3605	0.4134	UMUTeam	0.3547	0.3884
IIIA	0.6159	0.7932	UC3M-LCPM	0.3419	0.3856			
UC3M-LCPM	0.6093	0.7906						
UMUTeam	0.5955	0.7813						
VerbaNex-AI	0.4990	0.7279						
Baseline max freq.	0.2967	0.4604	Baseline max freq.	0.2205	0.2302	Baseline max-freq.	0.1881	0.1151
Baseline xlm-roberta-large	0.6575	0.8184	Baseline xlm-roberta-large	0.5406	0.5569	Baseline xlm-roberta-large	0.4769	0.4067

Table 2: Results of subtasks 1A, 1B, and 1C for tweets in Spanish (best run).

was narrower than in Subtask 1B, with the best English run surpassing the best Spanish run by 4 F1 points.

3 Task 2: Automatic Detection of Narratives

This task is posed as a multi-class, multi-label classification problem. Given a set of predefined narratives from each international actor (see Section 3.1), systems must determine to which narrative the tweets in the test set belong. The systems will receive the description of each narrative and a few examples of tweets in both languages (English and Spanish) that belong to each narrative (few-shot). A tweet can be associated with one, several or none of the reference narratives.

3.1 Strategic narratives

Our selection of target narratives is based on previous comprehensive analyses conducted by our team (Moral, 2024; Moral, 2023; Moral and Marco, 2023). The narratives are categorized based on the originating international actors. For each actor, we include 6 possible narratives. We aim to suggest overarching narratives that are broad enough to be generalizable in other contexts and are likely to incorporate a wide lexical variety of messages, thus reducing potential bias of the systems. The statements shown in Table 4 are based on plausible outcomes suggested by each of the narratives. A detailed description with examples of each narrative can be found at our website.[4]

3.2 Dataset

A newly created dataset based on the one provided in Task 1 has been used. We have two different datasets depending on the language of the diplomats' tweets. These sets have been divided into two files, the training set and the test set.

Each tweet has six narratives associated with it, depending on the country or region of the diplomat who wrote the tweet. Within each of these narratives there can be 3 tags:

- **Yes**, when the reading of the tweet is clearly in favour of the narrative. It is one of its main communicative intentions.
- **Leaning**, despite the narrative is not a primary communicative intention, there may be some reading of the tweet supporting the narrative. In other words, the narrative could be a secondary communicative intention.

[4]https://sites.google.com/view/dipromats2024/task-2/set-of-narratives

Both languages								
Subtask 1A			**Subtask 1B**			**Subtask 1C**		
Team	**ICM**	**F1**	**Team**	**ICM**	**F1**	**Team**	**ICM**	**F1**
Gold	1	1	Gold	1	1	Gold	1	1
Victor Vectors	**0.6619**	**0.8169**	**DSHacker**	**0.4530**	**0.6029**	**DSHacker**	**0.4945**	0.4611
DSHacker	0.6596	0.8151	Victor Vectors	0.4682	0.5868	Victor Vectors	0.4274	**0.4795**
PropaLTL	0.6318	0.8002	UMUTeam	0.4378	0.55	UMUTeam	0.4122	0.4476
UMUTeam	0.6302	0.7992	UC3M-LCPM	0.3353	0.4486			
UC3M-LCPM	0.6003	0.7813						
IIIA	0.5890	0.7736						
VerbaNex-AI	0.2413	0.4694						
Baseline max freq.	0.3011	0.4568	Baseline max freq.	0.2165	0.2284	Baseline max-freq.	0.1757	0.1142
Baseline xlm-roberta-large	0.6402	0.8049	Baseline xlm-roberta-large	0.5555	0.6468	Baseline xlm-roberta-large	0.5079	0.4972

Table 3: Results of subtasks 1A, 1B and 1C for tweets in Spanish and English (best run).

- **No**, when the tweet is completely unrelated to the narrative, or doesn't support it in any reading.

The choice of these three labels has been made to facilitate the labelling task. In Table 5 you can see the distribution of the labels per narrative in the test and training sets in the two languages. However, as the classification system for each tweet and narrative has to be binary, there will be two main evaluation measures depending on how the Leaning cases are considered.

3.3 Evaluation measures

For each narrative i with $i \in \{1, \cdots, 24\}$, we have the following confusion matrix shown in the Table 6

We have two evaluation measures, strict F_1 and lenient F_1. For each narrative, the following values will be calculated: the F_1 score, the $Macro\text{-}F_1$ score, and the $Micro\text{-}F_1$ score. For these measures we will use the formula:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

where precision is denoted by P, and recall by R.

Strict F_1

Measuring systems performance for the identification of tweets with narratives as primary communicative intention. Strict F_1 will be calculated from the precision and recall **over class Yes** for each narrative. In other words, the Leaning cases in the gold-standard will be considered as class No.

In this way, for each narrative i:

$$P_i = \frac{tp_i}{tp_i + fp_i + yl_i} \qquad R_i = \frac{tp_i}{tp_i + fn_i}$$

therefore,

$$F_{1i} = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i}.$$

With these F_{1i} values, we calculate

$$Macro\text{-}F_1 = \frac{1}{24} \sum_{i=1}^{24} F_{1i}.$$

To calculate $Micro\text{-}F_1$ value we have

$$P_{\text{micro}} = \frac{\sum_{i=1}^{24} tp_i}{\sum_{i=1}^{24} (tp_i + fp_i + yl_i)}$$

China (CH)	**Russia (RU)**	**European Union (EU)**	**United States (USA)**
CH1: The West is immoral, hostile and decadent.	RU1: The West and their allies are immoral, hostile and decadent.	EU1: The European Union is a successful story.	US1: The US has an honorable tradition.
CH2: China is a benevolent power.	RU2: Russia leads an alternative system to that sponsored by the West.	EU2: The European Union is united.	US2: The US has an admirable society.
CH3: China has an epic history.	RU3: Russia is a benevolent partner.	EU3: The European Union is useful for its citizens.	US3: The US has hostile enemies.
CH4: China has an appropriate political system and honorable values.	RU4: Russian history is admirable.	EU4: The European Union is an avant-garde political actor.	US4: The US is a force for good.
CH5: The government of the Chinese Communist Party succeeds.	RU5: Russia is at the world's forefront.	EU5: The European Union a global champion of fair causes.	US5: The US is a great international ally.
CH6: China is an interesting country in terms of culture, nature and heritage.	RU6: Russia is an interesting country in terms of culture, nature and heritage.	EU6: The European Union contributes to a better world.	US6: The US is at the forefront.

Table 4: Title of the narratives from each of the regions.

and

$$R_{\text{micro}} = \frac{\sum_{i=1}^{24} tp_i}{\sum_{i=1}^{24} (tp_i + fn_i)}$$

then,

$$Micro\text{-}F_1 = 2 \cdot \frac{P_{\text{micro}} \cdot R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}.$$

Lenient F_1

Measuring systems performance for the identification of tweets with narratives as primary or secondary communicative intention. Lenient F_1 will be calculated from the precision and recall **over classes Yes and Leaning** for each narrative. Cases labelled as Leaning in the gold-standard will be considered as YES or NO in the way that better fits with the model under evaluation. In this way, it can be considered as an upper bound of systems performance.

In this way, for each narrative i:

$$P_i = \frac{tp_i + yl_i}{tp_i + yl_i + fp_i} \qquad R_i = \frac{tp_i + yl_i}{tp_i + yl_i + fn_i}$$

therefore,

$$F_{1i} = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i}.$$

With these F_{1i} values, we calculate

$$Macro\text{-}F_1 = \frac{1}{24} \sum_{i=1}^{24} F_{1i}.$$

To calculate $Micro\text{-}F_1$ value we have

$$P_{\text{micro}} = \frac{\sum_{i=1}^{24} (tp_i + yl_i)}{\sum_{i=1}^{24} (tp_i + yl_i + fp_i)}$$

	Spanish dataset						English dataset					
	Train set			Test set			Train set			Test set		
	yes	**l**	**no**	**yes**	**l**	**no**	**yes**	**l**	**no**	**yes**	**l**	**no**
CH1	2	0	11	21	8	171	4	1	10	38	34	128
CH2	2	2	9	65	9	126	5	2	8	39	50	111
CH3	2	2	9	6	0	194	3	1	11	8	11	181
CH4	2	0	11	24	24	152	4	1	10	14	70	116
CH5	2	2	9	12	21	167	3	1	11	7	31	162
CH6	4	0	9	22	11	167	2	0	13	13	18	169
China	**14**	**6**	**58**	**150**	**73**	**977**	**21**	**6**	**63**	**119**	**214**	**867**
RU1	2	1	9	34	12	154	2	1	10	30	21	149
RU2	2	1	9	4	15	181	2	1	10	14	23	163
RU3	2	0	10	30	28	142	4	0	9	52	37	111
RU4	2	1	9	25	19	156	3	0	10	25	13	162
RU5	2	0	10	22	25	153	2	0	11	14	26	160
RU6	2	2	8	14	10	176	2	1	10	27	7	166
Russia	**12**	**5**	**55**	**129**	**109**	**962**	**15**	**3**	**60**	**162**	**127**	**911**
EU1	2	1	9	14	10	176	2	3	9	0	28	172
EU2	2	0	10	29	21	150	2	0	12	0	32	168
EU3	2	2	8	24	23	153	2	3	9	10	19	171
EU4	2	2	8	31	23	146	2	2	10	21	37	142
EU5	2	1	9	79	33	88	3	0	11	20	59	121
EU6	4	0	8	48	65	87	3	2	9	48	87	65
European Union	**14**	**6**	**52**	**225**	**175**	**800**	**14**	**10**	**60**	**99**	**262**	**839**
US1	3	4	6	23	1	175	2	2	8	10	17	173
US2	2	1	10	19	5	175	2	1	9	23	30	147
US3	2	1	10	46	11	142	2	1	9	32	24	144
US4	4	2	7	66	23	110	3	1	8	50	52	98
US5	2	0	11	37	18	144	2	1	9	45	28	127
US6	2	0	11	11	4	184	2	1	9	22	23	155
USA	**15**	**8**	**55**	**202**	**62**	**930**	**13**	**7**	**52**	**182**	**174**	**844**

Table 5: Distribution of labels by narratives in training and test sets.

		Actual values		
		yes	**leaning**	**no**
Predicted values	**yes**	tp_i	yl_i	fp_i
	no	fn_i	nl_i	tn_i

Table 6: Confusion matrix for each narrative.

and

$$R_{\text{micro}} = \frac{\sum_{i=1}^{24}(tp_i + yl_i)}{\sum_{i=1}^{24}(tp_i + yl_i + fn_i)}$$

then,

$$Micro\text{-}F_1 = 2 \cdot \frac{P_{\text{micro}} \cdot R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}.$$

3.4 System results

Table 7 shows the results of the $Micro - F1$ metric for the two languages, for all teams and for the baseline model.

Two teams have been submitted for this task. One of them, *UMUTeam*, has submitted two runs and the *Mancha_Azul* team only one run.

The umuteam approaches are based on

	Spanish			English		
	F1 Strict	**F1 Lenient**	**F1 Avg**	**F1 Strict**	**F1 Lenient**	**F1 Avg**
Mancha_Azul	**0.5643**	**0.7178**	**0.6411**	**0.4831**	**0.7390**	**0.6111**
Baseline	0.3769	0.5278	0.4524	0.2875	0.5446	0.4161
umuteam-Zephyr	0.3046	0.4427	0.3737	0.3149	0.5265	0.4207
umuteam-TuLu	0.2729	0.3976	0.3353	0.2303	0.4441	0.3372

Table 7: Evaluation results on the test set of all participants.

a Few-Shot learning strategy using TuLu and Zephyr models. In contrast, the Mancha_Azul approach is a multi-agent signal-based LLM system.

In addition, the results of a baseline model are included. The baseline model is a Mixtral 8x7B Instruct (Jiang et al., 2024) with a 4-bit quantization (Jacob et al., 2017). For this model, a Zero-Shot strategy has been employed.

Mancha_Azul achieved the best measure of F_1 in the two languages with values of F_1 *Lenient* of 0.7178 for Spanish and 0.7390 for English. This team achieves a significant improvement over the baseline model of 36% for Spanish and 36.7% for English. However, we can see that the models proposed by the *UMUTeam* team have a slightly lower performance than the baseline model in almost all measures.

4 Conclusions

This paper presented an overview of the second edition of DIPROMATS, a shared task at IberLEF 2024, which challenged participants to automatically categorize propaganda and strategic narratives in tweets from diplomatic authorities. This edition featured two tasks: the continuation of DIPROMATS 2023 and a new challenge focusing on narrative detection.

The first task introduced a revised typology of techniques, resulting in a more balanced dataset compared to the previous edition. Building on the efforts of last year's participants, the most successful approaches for Task 1 utilized BERT-based models with various optimization and data augmentation strategies. Notably, the results of this edition have improved significantly, with the exception of the best English run in Subtask 1C, which scored lower than last year's. In particular, performance in Spanish saw considerable enhancement, narrowing the gap between the two languages. This suggests that the refinement of participant strategies and a less complex dataset were effective.

Although participation in Task 2 was low, it remains valuable for several reasons. Firstly, to our knowledge, this was the first attempt to incorporate the concept of strategic narratives into an NLP challenge, thus maintaining DIPROMATS as a significant initiative for integrating international political communication into computational tasks. Secondly, Task 2 served as an exploratory experiment to assess the feasibility of identifying narratives without relying solely on lexical similarity. The results suggest that this is plausible, and the introduction of an intermediate category, the "leaning" class, has proven to be effective both qualitatively and in terms of system performance.

All in all, both tasks of this challenge have once again demonstrated that the automated detection of propaganda and narratives are complex tasks with significant room for improvement. The two editions of DIPROMATS have successfully advanced the study of these topics and provided valuable insights on the path forward. This challenge emphasizes the potential for automated systems to assist in identifying manipulative content, contributing to the fight against hostile information by transferring knowledge from communication studies to practical applications in technology.

Acknowledgments

This work was supported by DeepInfo project (AEI PID2021-127777OB-C22) and HAMiSoN project grant CHIST-ERA-21-OSNEM-002, AEI PCI2022-135026-2 (MCIN/AEI/10.13039/501100011033 and EU "NextGenerationEU"/PRTR).This work was also 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 "Misinformation and Miscommunication in social media: bias (MISMIS-BIAS)" under grant PGC2018-096212-B-C32. 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 (2024–2025). 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.

References

Alisetti, S. V. 2024. Paraphrase Generator with T5, June.

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.

Bolsover, G. and P. Howard. 2017. Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda. *Big Data*, 5(4):273–276, December.

Bolt, N. 2012. *The violent image: insurgent propaganda and the new revolutionaries*. Hurst & Company, London. OCLC: 1233055622.

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.

Colley, T. 2020. Strategic narratives and war propaganda. In P. Baines, N. O'Shaughnessy, and N. Snow, editors, *The SAGE handbook of propaganda*. SAGE, London, pages 491–508.

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., 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.

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.

García-Díaz, J. A., P. J. Vivancos-Vicente, Á. Almela, and R. Valencia-García. 2022. UMUTextStats: A linguistic feature extraction tool for Spanish. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6035–6044, Marseille, France, June. European Language Resources Association.

Jacob, B., S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. 2017. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference, December.

Jiang, A. Q., A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. 2024. Mixtral of Experts, January.

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].

Miskimmon, A., B. O'Loughlin, and L. Roselle. 2013. *Strategic narratives: communication power and the new world order*. Number 3 in Routledge studies in global information, politics and society. Routledge, Taylor & Francis Group, New York ; London.

Moral, P. 2023. Restoring reputation through digital diplomacy: the European Union's strategic narratives on Twitter during the COVID-19 pandemic. *Communication & Society*, pages 241–269, April.

Moral, P. 2024. A tale of heroes and villains: Russia's strategic narratives on twitter during the covid-19 pandemic. *Journal of Information Technology & Politics*, 21(2):146–165.

Moral, P. and G. Marco. 2023. Assembling stories tweet by tweet: strategic narratives from Chinese authorities on Twitter during the COVID-19 pandemic. *Communication Research and Practice*, 9(2):159–183, April.

Moral, P., G. Marco, J. Gonzalo, J. Carrillo-de Albornoz, and I. Gonzalo-Verdugo. 2023. 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, September.

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.

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.

Richards, J. 2023. The Use of Discourse Analysis in Propaganda Detection and Understanding. In *Routledge Handbook of Disinformation and National Security*. Routledge, London, 1 edition, October, pages 385–400.

Riessman, C. K. 2008. *Narrative methods for the human sciences*. Sage Publications, Los Angeles.

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

Zhang, X., Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, and A. El-Kishky. 2023. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations at Twitter, August. arXiv:2209.07562 [cs].