UniLeon-UniBO at IberLEF 2023 Task DIPROMATS: RoBERTa-based Models to Climb Up the Propaganda Tree in English and Spanish

Francisco Jáñez-Martino[1], Alberto Barrón-Cedeño[2]

[1]Department of Electrical, Systems and Automation, Universidad de León, León, Spain
[2]DIT, Università di Bologna, Forlì, Italy

Abstract

We describe our participation to the IberLEF 2023 shared task DIPROMATS on the automatic detection of propaganda in tweets posted by diplomats from different geographic regions in English and Spanish. Whereas Task 1 aims at detecting propaganda (binary task), Task 2 and Task 3 seek to categorize the type of propaganda in four groups and 15 techniques, as a multilabel classification problem. We design a pipeline to face all three tasks by employing four multi-label model —one for each group— in order to spot the 15 propaganda techniques and then, we climb up to identify their group and, finally, respond to the binary classification. Our official submission to the English tasks, built on top of RoBERTa, achieves an overall ICM-Hard score of 0.1835 for Task 1 (3rd position out of 30 submissions), 0.1342 for Task 2 (2nd position out of 28 submissions) and 0.0693 for Task 3 (5th position out of 30 submissions). Our official submission to the Spanish tasks, which is based on BERTIN, obtains 0.6301 for Task 1 (fourth position out of 18 submissions), -0.0134 for Task 2 (first position out of 17 submissions) and -0.1478 for Task 3 (first position out of 17 submissions).

Keywords

multilabel propaganda identification, Twitter, persuasion techniques identification, Diplomats bias

1. Introduction

According to [1] propaganda is "the expression of opinion or action by individuals or groups deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends". In contrast to misinformation or disinformation, propagandist content is not necessarily factually false and its intention becomes evident only through a careful and educated observation.

The DIPROMATS shared task at IberLEF 2023 [2], proposed the challenge of creating supervised models for the identification of specific propaganda techniques. They focus on the identification of propaganda spread by official diplomatic authorities on Twitter coming from China, Russia, the European Union, and the United States published both in Spanish and English. The challenge includes tweets from Twitter profiles that belong to governmental accounts, embassies, ambassadors, and other diplomatic profiles such as consuls and missions. The

IberLEF 2023, September 2023, Jaén, Spain

francisco.janez@unileon.es (F. Jáñez-Martino); a.barron@unibo.it (A. Barrón-Cedeño)
0000-0001-7665-6418 (F. Jáñez-Martino); 0000-0003-4719-3420 (A. Barrón-Cedeño)

DIPROMATS task adapts a list of techniques originally proposed by San Martino et al. [3] by incorporating techniques inspired by Aristotle's principles of rhetoric [4] and political purposefulness [5]. Table 1 show figures on the instances provided as training material in both English and Spanish, clustered into four propaganda super-groups. With this materials at hand, DIPROMATS proposed three tasks:

Task 1 Propaganda identification A system must decide whether a given tweet contains propaganda techniques. This is a binary classification problem.

Task 2 Propaganda coarse-grained characterization The categorization considers multiple techniques identified in the literature that are clustered according to their rhetorical features into four groups: appeal to commonality, discrediting the opponent, loaded language, and appeal to authority.

Systems have to decide, for each tweet, in which of the available four categories it fits (plus a negative class). The proposed typology can be found here [6]. Evaluation will consider a coarse-grain categorization with four classes of propaganda plus a negative class (Task 2), and a fine-grained categorization with 15 subclasses plus a negative class (Task 3).

Task 3 Propaganda fine-grained characterization The categorization considered 15 fine-grained persuasion techniques: ad populum, flag waving, absurdity appeal, demonization, doubt, fear appeals (destructive), name calling, personal attack, propaganda slinging, reductio ad Hitlerum, scapegoating, undiplomatic assertiveness/whataboutism, loaded language, appeal to false authority, and bandwagoning (plus a negative class). This is a multi-label classification task.

We follow a bottom-up strategy to address the three tasks, departing from Task 3, the finest-grained level. Given a tweet, we try to identify the specific technique among the 15 possible ones (Task 3). On the basis of this decision, we determine to which of the four super-group the tweet belongs to (Task 2). Finally, we perform the binary decision, whether the tweet is propaganda or not (Task 1).

The models for both English and Spanish are built on top of RoBERTa [7], which follows a Transformer architecture [8]. For Spanish, we use the RoBERTa architecture from the BERTIN project [9]. Other than the *standard* fine-tuning, we enhance it by applying a pre-processing stage to analyse Twitter properties such as hashtags and user mentions, as well as adding information from metadata. We also experiment with adding external materials from the Propaganda Techniques Corpus [3] to increase the number of instances for certain techniques.

Out bottom-up strategy results in the top performance on both Task 2 and Task 3 Spanish as well as competitive performance in the other four tasks (from second to fifth ranking).

The rest of the paper is structured as follows. Section 2 reviews related work on propaganda detection. Section 3 describes our baselines and the preprocessing strategies as well as the combination and tweaking of the transformer modules to perform the fine- and coarse-grained classification. Section 4 describes our experimental setup as well as the preliminary and official submission results. Finally, Section 5 closes with conclusions and future work.

	en	**es**
Binary Classification		
Propaganda	1,974	1,199
Non Propaganda	6,434	4,921
G1 Appeal to commonality	**617**	**234**
t1 ad populum	72	–
t2 flag waving	545	234
G2 Discrediting the opponent	**980**	**925**
t3 absurdity appeal	30	19
t4 demonization	44	41
t5 doubt	76	27
t6 fear appeals (destructive)	61	57
t7 name calling	213	90
t8 personal attack	-	-
t9 propaganda slinging	114	124
t10 reductio ad Hitlerum	-	-
t11 scapegoating	12	4
t12 undiplomatic assertiveness / whataboutism	430	563
G3 Loaded language	**913**	**389**
t13 loaded language	913	389
G4 Appeal to authority	**4**	**6**
t14 appeal to false authority	2	6
t15 bandwagoning	2	–

Table 1: Number of propagandist and non-propagandist instances (Task 1, top), per coarse-grained propaganda group (Task 2, bolded), and per fine-grained technique (Task 3) in English and Spanish.

2. Related Work

There has been a rise in the interest for the development of models for spotting propaganda [10]. The growing use of social media and online websites to stay informed poses a challenge for the detection of new forms of propaganda [11]. This biased information is used in messages to influence communities and agendas [12].

Several competitions have been held over the last few years that have fostered the development of technology for propaganda identification [3, 13, 14, 15].

Some approaches have focused on the analysis at the document level. Barrón-Cedeño et al. [16] evaluated different sets of manually-engineered features along with a Maximum Entropy classifier to detect the level of propaganda in outlet news. On the same way, Barfar [17] applied SVM, (shallow/deep) neural networks, and LightGBM taking a hybrid linguistic/game-theoretic approach.

Other approaches have looked into finer-grained levels, such as sentences or tweets [18, 19].

Orlok et al. [18] proposed an unsupervised approach using user behaviours and text analysis. Da San Martino et al. [3] designed a competition to spot specific propaganda techniques. Most of the top models were based on attention models (e.g., [8]). Vorakitphan el al. [20] combined the outputs from sentence-span based RoBERTa with the feature-based BiLSTM to detect propaganda snippets from plain text.

Most of high-performance models are based on RoBERTa architecture. Thus, we depart from this architecture to develop our systems.

3. Methodology

We develop a system that concurrently performs all tasks and generates the three corresponding outputs. Figure 1 shows an overview of our pipeline.

Preprocessing Following Pota et al. [21], we run into the Twitter-specific features and how they could be useful for the task at hand. We keep emoticons, emoji, emails, phone numbers and dates, among others. Both hashtags and user mentions are split into words using wordninja [22]. The out-of-the-box model is intended for English, thereby, we use it directly. For Spanish, we trained a model from scratch, using a list with 10,000 of the most frequent words according to Wiktionary [23, 24]. We remove URLs.

Fine-grained classification We start by spotting specific propaganda techniques. For that, we fine-tune four models based on the RoBERTa architecture, one for each of the propaganda groups. The outputs for each of the RoBERTa models is a combination of fine- and coarse-grained classes: the techniques within its own group and the three other groups. For example, the outputs for $Roberta_1$ are *Ad populum / Ad antiquitatem* and *Flag Waving* along with *Group 2*, *Group 3* and *Group 4*. The fine-grained decisions t_j of each of the four modules represent the output for Task 3 and are fed to the group-decision stage. The coarse-grained decisions are discarded.

Coarse-grained classification Having the fine-grained classification as input, we climb up to infer the group. This is simply an OR logic gate, which gets activated if any the techniques within that group has been spotted.

Binary classification We follow an identical approach as for the coarse-grained classification, this time with one single OR logic gate having as input the four groups.

Contextualization and data augmentation We noticed a trend in the posts for tweets that were labeled as propagandist and included mentions of the country of origin of the poster. In order to assess the volume of this trend, we computed the number of tweets that mention the country of origin.

Table 3 shows statistics on that respect. We use a list of international actors, acronyms and demonym variations to match these occurrences; see Appendix A for details. Whereas

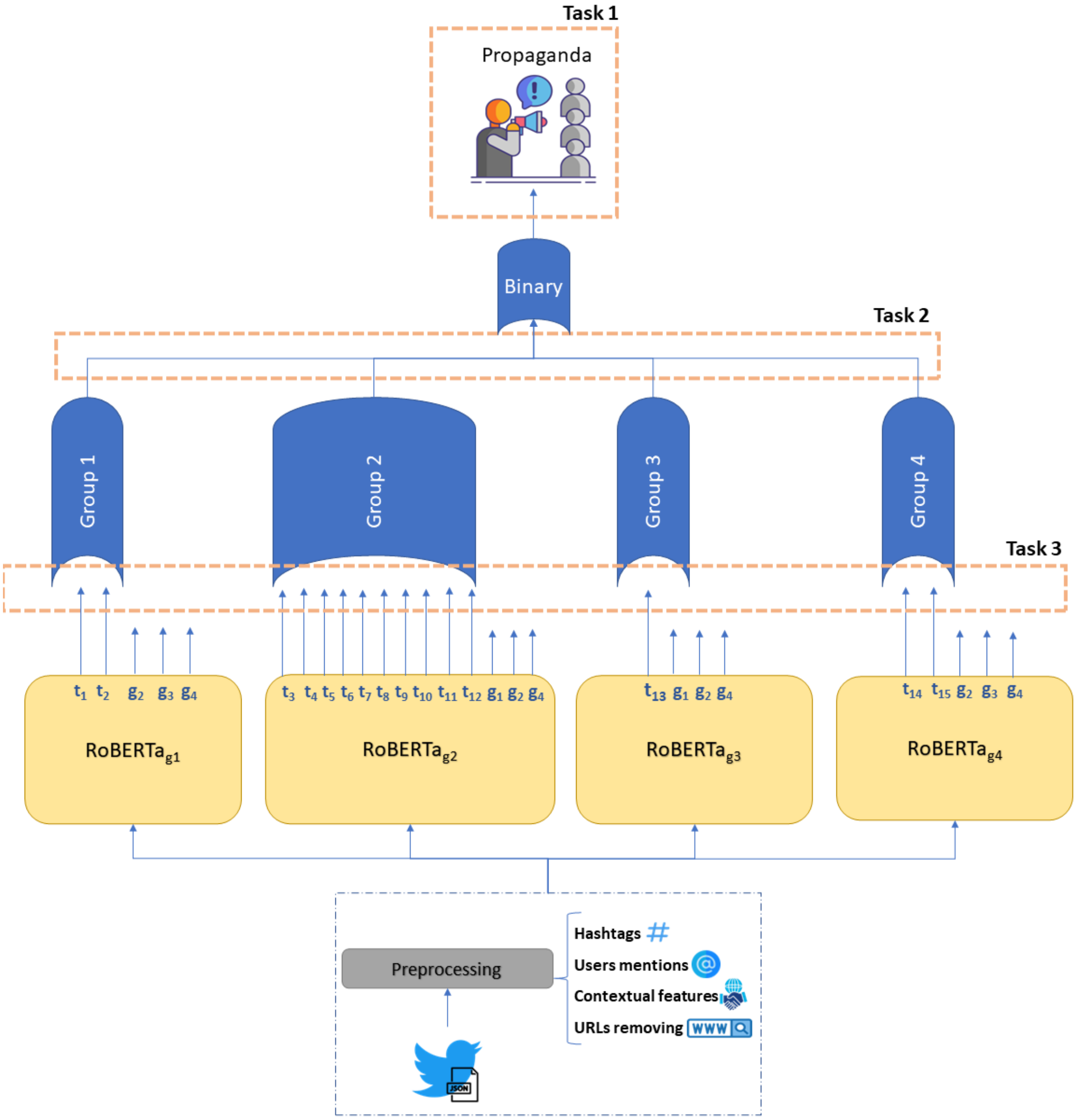

Figure 1: Representation of our top-proposed pipeline; g_i represents each of the coarse-grained propaganda groups (Task 2; $i \in [1, 4]$), t_j represents each of the fine-grained propaganda techniques (Task 3; $j \in [1, 15]$).

between 40 and 55% of the propagandist tweets from Chinese (Russian) accounts do mention China (Russia), less than 25% of the non-propagandist tweets mention it. This trend is not as differentiated in tweets from the European Union since, not being one single country, European parties hardly mention the EU in their posts, which may be explained since EU is a supranational political and economic union of 27 member states and their diplomats may name their countries of origin. Still, propagandist tweets tend to mention the EU slightly more. The accounts with

	China		EU		Russia		USA	
	en	**es**	**en**	**es**	**en**	**es**	**en**	**es**
Non-propagandist								
Total	1,486	1,613	1,859	1,429	1,600	598	1,489	1,281
Naming the country	304	234	205	81	412	102	1,100	1,129
Percentage (%)	(20.46)	(14.51)	(11.03)	(5.67)	(25.75)	(17.06)	(73.88)	(88.13)
Propagandist								
Total	684	565	184	79	405	197	701	358
Naming the country	351	224	21	6	160	109	303	154
Percentage (%)	(51.32)	(39.65)	(11.41)	(7.59)	(39.51)	(55.33)	(43.22)	(43.02)

Table 2: Total number and percentage (%) of propagandist and non-propagandist tweets when the poster mentions her own international actor in both English and Spanish.

origin in the US are an exception, since they tend to mention their country more frequently and do so more often in non-propagandist posts.

In order to make the model aware of this information, we inject it as a contextual feature. We attach the sentence "This has been written from [country]." in English and "Este tweet ha sido escrito desde [country]." in Spanish at the beginning of all tweets before feeding them to the model.

As for data augmentation, we consider the Propaganda Techniques Corpus (PTC), developed in the framework of SemEval 2020 Corpus, Da San Martino et al. [3] PTC is composed of a set of news articles written in English from a period between mid-2017 and early 2019. The dataset includes a total of 8,981 propagandist snippets or diverse lengths. In order to increase the volume of the training material, we selected and extracted the sentences from PTC that contained the same (or similar) techniques to those considered in DIPROMATS.

4. Experimentation

Experimentation Setup. We used roberta-large model [25] and bertin-roberta-base-spanish [26] for English and Spanish, respectively. We trained every model during 20 epochs with 16 as the batch size for all cases except for the case in which the PTC corpus was added to the training set that was trained during 5 epochs. We randomly selected the 20% of the training instances to held out for validation purposes. We carried out the experiments on computer with 128GB of RAM, two processor Intel Xeon E5-2630v3 of 2,4GHz and two Nvidia Titan X.

In addition to standard classification metric F1-Score, DIPROMATS also reported the metric ICM [27], which is the official metric of the task.

Internal Experiments. Table 4 show the results of the different models on the validation set, including the vanilla baseline and the inclusion or not of a pre-processing stage or external training material (only for English).

For English, the baseline alone consistently performs worst in all three tasks. The inclusion of the pre-processing boosts the performance. Still, the further addition of external material has a slight negative impact on the outcome. In terms of ICM metric, Spanish baseline with

	English					
	Task 1		Task 2		Task 3	
	ICM	F1	ICM	F1	ICM	F1
Baseline	0.2813	0.8159	0.0309	0.4787	-0.0365	0.4708
∟ preprocessing	0.2973	0.8219	0.0593	0.4893	0.0061	0.4878
∟ PTC	0.2920	0.8201	0.0444	0.4818	-0.0434	0.4490
	Spanish					
	Task 1		Task 2		Task 3	
	ICM	F1	ICM	F1	ICM	F1
Baseline	-0.1466	0.6098	-0.6802	0.1301	-0.9573	0.0452
∟ preprocessing	-0.1776	0.5551	-0.6391	0.1223	-0.8955	0.0956

Table 3: Performance on the validation set, during the internal experimentation with the different alternatives.

preprocessing stage also obtained the highest performance in Task 2 and Task 3, however, its performance was poorer in Task 1.

Official Submissions We submitted all five system alternatives : three for English and two for Spanish. Table 4 shows the performance obtained by the top-5 official submissions to the task. For all tasks, only the top-performing of our systems is included, which is aligned with the results on development (i.e. the baseline+preprocessing variant both for English and for Spanish. Our submission for English reached the third, second and fifth positions overall Task 1, Task 2 and Task 3, respectively. Our Spanish submission reachved the fourth, first and first positions in Task 1, Task 2 and Task 3, respectively. These results validated the positive impact of the preprocessing stage in our models as it we could see in our internal experiments.

5. Conclusions and Future Work

We described our bottom-up approach address the propaganda identification tasks proposed at the 2023 edition of DIPROMATS at IberLEF. We addressed the task at multiple granularities by plugging logical gates to the combination of the output of different parallel RoBERTa-based neural architectures. Our strategy first tries to identify among the 15 fine-grained labels and climbs up through the four coarse-grained decisions up to making a binary decision: propaganda or not. Our experiments showed that performing Twitter-specific pre-processing and adding contextual source information is crucial to improve the performance and that the inclusion of similar (our-of-genre) datasets annotated for propaganda do not seem to help. Our approach performed satisfactorily on the shared task, reaching two first positions (Spanish Task 2 and Task 3), one second (English Task 2) and one third position (English Task 1) overall.

As future work, we will investigate how to combine the coarse-grained classification decisions (currently neglected and substituted by a Boolean decision) to give a more robust output regarding group detection. We are also assessing the potential of using an ensemble of all fine-grained models to produce feature vectors to feed to a machine learning classifier.

English								
Task 1			**Task 2**			**Task 3**		
Team	ICM	F1		ICM	F1	Team	ICM	F1
Lian Tian	0.2013	0.6784	Albert Pritzkau	0.1778	0.5591	Albert Pritzkau	0.1227	0.4838
PropaLTL	0.1957	0.6777	**UniLeon -UniBO**	**0.1342**	**0.549**	Albert Pritzkau	0.1018	0.4824
UniLeon -UniBO	**0.1835**	**0.6667**	Albert Pritzkau	0.1299	0.5465	Albert Pritzkau	0.0865	0.4645
PropaLTL	0.1817	0.6667	Albert Pritzkau	0.0955	0.5395	Albert Pritzkau	0.0794	0.4715
PropaLTL	0.1793	0.6594	Albert Pritzkau	0.0913	0.5149	**UniLeon -UniBO**	**0.0693**	**0.4405**
Spanish								
Task 1			**Task 2**			**Task 3**		
Team	ICM	F1		ICM	F1	Team	ICM	F1
PropaLTL	0.1724	0.6681	**UniLeon -UniBO**	**-0.0134**	**0.4301**	**UniLeon -UniBO**	**-0.1478**	**0.2788**
umuteam	0.1323	0.631	umuteam	-0.0180	0.4164	VRAIN-ELiRF	-0.1576	0.3628
umuteam	0.1316	0.6301	umuteam	-0.0192	0.416	VRAIN-ELiRF	-0.1694	0.3884
UniLeon -UniBO	**0.1254**	**0.6301**	VRAIN-ELiRF	-0.0369	0.4578	VRAIN-ELiRF	-0.1780	0.3943
PropaLTL	0.1141	0.6105	VRAIN-ELiRF	-0.0379	0.4626	umuteam	-0.1810	0.3414

Table 4: Extract of the top-5 official submissions to the DIPROMATS shared task in English and Spanish. Submissions ranked on the basis of ICM-Hard (ICM) terms.

References

[1] I. for Propaganda Analysis (Ed.), Proppy: Organizing the news based on their propagandistic content, Propaganda analysis. volume i of the publications of the institute for propaganda analysis (1938) 210–218.

[2] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de-Albornoz, Iván 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).

[3] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1377–1414. doi:10.18653/v1/2020.semeval-1.186.

[4] C. Miles, P. Baines, N. O'Shaughnessy, N. Snow, Rhetorical methods and metaphor in viral propaganda, The SAGE Handbook of Propaganda. SAGE. 244-260 (2019) 154–170.

[5] K. Johnson-Cartee, G. Copeland, Strategic political communication: Rethinking social influence, persuasion, and propaganda, Rowman & Littlefield Publishers (2004).

[6] IberLEF, DIPROMATS 2023 - Background, https://sites.google.com/view/dipromats2023/background, year=2023, note = Accessed: June 2023, ????

[7] 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.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[9] J. D. la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23.

[10] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, P. Nakov, A survey on computational propaganda detection, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20, 2021, p. 7.

[11] M. Hashemi, M. Hall, Detecting and classifying online dark visual propaganda, Image and Vision Computing 89 (2019) 95–105. URL: https://www.sciencedirect.com/science/article/pii/S0262885619300848. doi:https://doi.org/10.1016/j.imavis.2019.06.001.

[12] Y. Mejova, M. Petrocchi, C. Scarton, Special issue on disinformation, hoaxes and propaganda within online social networks and media, Online Social Networks and Media 23 (2021) 100132. URL: https://www.sciencedirect.com/science/article/pii/S2468696421000161. doi:https://doi.org/10.1016/j.osnem.2021.100132.

[13] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 70–98. doi:10.18653/v1/2021.semeval-1.7.

[14] F. Alam, H. Mubarak, W. Zaghouani, G. Da San Martino, P. Nakov, Overview of the WANLP 2022 shared task on propaganda detection in Arabic, in: Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 108–118.

[15] 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, volume 23, 2023.

[16] 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. doi:https://doi.org/10.1016/j.ipm.2019.03.005.

[17] A. Barfar, A linguistic/game-theoretic approach to detection/explanation of propaganda, Expert Systems with Applications 189 (2022) 116069. doi:https://doi.org/10.1016/j.eswa.2021.116069.

[18] M. Orlov, M. Litvak, Using behavior and text analysis to detect propagandists and misinformers on twitter, in: Conference: The 5th International Conference on Information Management and Big Data (SIMBig 2018), Track on Social Network and Media Analysis and Mining (SNMAM), 2018, pp. 1–8. doi:10.1007/978-3-030-11680-4_8.

[19] 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, WebSci '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 191–201. doi:10.1145/3501247.3531543.

[20] V. Vorakitphan, E. Cabrio, S. Villata, PROTECT: A Pipeline for Propaganda Detection and Classification, in: CLiC-it 2021- Italian Conference on Computational Linguistics, Milan, Italy, 2022, pp. 1–6. URL: https://hal.science/hal-03417019.

[21] M. Pota, M. Ventura, H. Fujita, M. Esposito, Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets, Expert Systems with Applications 181 (2021) 115119. doi:https://doi.org/10.1016/j.eswa.2021.115119.

[22] Keredson, Python package - wordninja, 2023. https://pypi.org/project/wordninja/, Accessed: June 2023.

[23] Matthias Buchmeier, Matthias buchmeier/spanish frequency list-5001-10000, 2023. https://en.wiktionary.org/wiki/User:Matthias_Buchmeier/Spanish_frequency_list-1-5000, Accessed: June 2023.

[24] Matthias Buchmeier, Matthias buchmeier/spanish frequency list-5001-10000, 2023. https://en.wiktionary.org/wiki/User:Matthias_Buchmeier/Spanish_frequency_list-5001-10000, Accessed: June 2023.

[25] Hugging Face, Hugging face roberta-large, 2023. https://huggingface.co/roberta-large, Accessed: June 2023.

[26] World Health Organisation, bertin-project/bertin-roberta-base-spanish · hugging face, 2023. https://huggingface.co/bertin-project/bertin-roberta-base-spanish, Accessed: June 2023.

[27] 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), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

A. Country-Mentioning Keywords

Table A.1 shows the list of terms that we consider to determine whether a diplomat is referring to her own international actor in a tweet.

International actor	List of terms	
	en	es
China	’China’, ’Chinese’	’China’, ’chinos’, ’chinas’
Russia	’Russia’, ’Russian’	’Rusia’, ’rusos’, ’rusas’
European Union	’European Union’, ’EU’, ’European’	’Unión Europea’, ’EU’,’europeos’, ’europeas’
USA	’USA’, ’United States’, ’U.S.’, ’US’, ’US-American’, ’United-Statesian’,	’Estados Unidos’,’EEUU’, ’EE.UU.’ ’estadounidenses’

Table A.1: Terms used to match tweets where the diplomat name its international actor.