PropaLTL at DIPROMATS: Incorporating Contextual Features with BERT's Auxiliary Input for Propaganda Detection on Tweets

Marco Casavantes[1,*], Manuel Montes-y-Gómez[1], Delia Irazú Hernández-Farías[1], Luis Carlos González[2] and Alberto Barrón-Cedeño[3]

[1]*Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico*
[2]*Universidad Autónoma de Chihuahua, Chihuahua, Mexico*
[3]*Università di Bologna, Forlì, Italy*

Abstract

In this paper, we describe our participation in the *Automatic Detection and Characterization of Propaganda Techniques from Diplomats* shared task (DIPROMATS). For this, we studied the inclusion of features grounded on contextual information. That is, our hypothesis was inspired in combining the text of tweets with contextual attributes such as its geographical origin, type of message, and emotions, aiming to improve the performance of the detection of propagandistic tweets. Our strategy of leveraging context data obtained very competitive scores in the *binary propaganda identification task*: the top position for Spanish out of 18 runs, and the second position for English out of 30 runs.

Keywords

Propaganda detection, contextual information, BERTweet, RoBERTuito

1. Introduction

Propaganda can be defined as "an evolving set of techniques and mechanisms which facilitate the propagation of ideas and actions" [1]. This phenomenon is associated with mainstream news outlets and political campaigns, such as newspapers or websites that provide news as their primary content. Social media has evolved over time, from representing a source of amusement to acting as the primary source of news for many individuals. As a result, they have become a preferred platform for the distribution of propaganda. Detecting propaganda in news and social media, in contrast to disinformation, has received little attention from journalists, fact-checkers, and scholars [2]. In light of this, DIPROMATS aims to promote research on this issue.

The DIPROMATS shared task at IberLEF 2023 proposes two challenges: (i) *Binary propaganda identification*, which involves deciding whether or not a tweet contains propagandist content,

IberLEF 2023, September 2023, Jaén, Spain
*Corresponding author.
mcasavantes@inaoep.mx (M. Casavantes); mmontesg@inaoep.mx (M. Montes-y-Gómez); dirazuhf@inaoep.mx (D. I. Hernández-Farías); lcgonzalez@uach.mx (L. C. González); a.barron@unibo.it (A. Barrón-Cedeño)
0000-0003-2339-2361 (M. Casavantes); 0000-0002-7601-501X (M. Montes-y-Gómez); 0000-0003-2133-8716 (D. I. Hernández-Farías); 0000-0003-1546-9752 (L. C. González); 0000-0003-4719-3420 (A. Barrón-Cedeño)

and (ii) *multiclass propaganda characterization*, which aims to categorize the type of propaganda among four classes (*"Appeal to Commonality"*, *"Discrediting the Opponent"*, *"Loaded Language"*, and *"Appeal to Authority"*) and a fine-grained categorization with 15 subclasses [3]. One of the most intriguing parts of this shared task is that the identification of propaganda on Twitter is focused on official authorities, such as government accounts, embassies, and other diplomatic entities.

The strategy of Team PropaLTL focused primarily on the binary propaganda identification task. The following insights served as inspiration for our approach:

(a) Transformers and pre-trained language models have shown a very competitive performance in a wide variety of natural language processing tasks, including the detection of propaganda [4, 5, 6]. The defined framework of DIPROMATS presents a perfect opportunity to continue assessing the effectiveness of these models in more particular situations.

(b) Contextual information has been successfully exploited in several tasks related to the classification of social media content, such as the detection of cyberbulling [7], aggressive contents [8, 9], abusive content [10], political intent [11], and also propaganda detection [6]. The DIPROMATS datasets include some contextual information that can be potentially useful for addressing the task in hand.

Our intuition is that propaganda can manifest differently depending on its context. Accordingly, we decided to exploit both transformers-based models along with data regarding the context of tweets. In particular, we considered three additional aspects of a given tweet: (i) The country of origin of its author, (ii) the way in which it has been disseminated (i.e., original tweet, retweet, quote or reply), and (iii) the most likely emotion that it evokes. Whereas the first two aspects are metadata included in the dataset, emotion was inferred with a supervised model. When designing our approach, we took advantage of the possibility offered by BERT models to introduce sequences of tokens packed together with an instance by means of an auxiliary input. Specifically, we used this auxiliary input to include the three contextual attributes described above.

The rest of this paper is structured as follows. Section 2 describes the proposed approach. Section 3 covers our experimental settings. Section 4 discusses the obtained results. Finally, Section 5 draws our conclusions about our participation in the shared task.

2. Description of the Proposed Approach

2.1. Task 1: Binary propaganda identification

Figure 1 outlines our approach to Task 1. It consists of three main modules: data preparation, features computation, and BERT-based classification.

Data preparation. We extracted the country of the tweet —that from the user who posted it— and the mechanism by which the tweet was disseminated. The text of each tweet, from

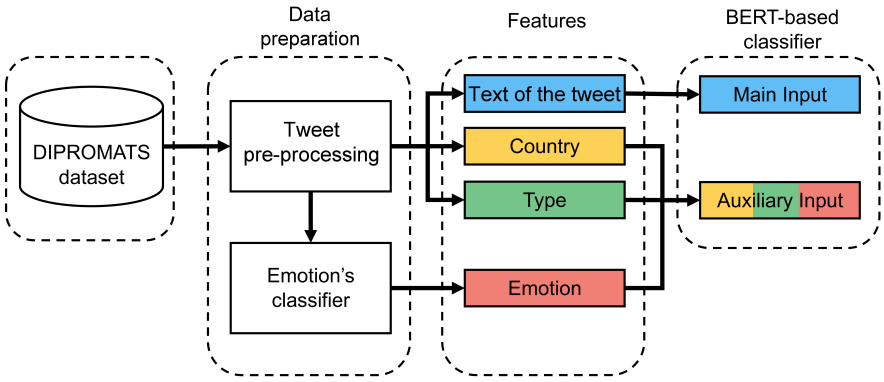

Figure 1: Representation of the pipeline for Task 1.

now on referred to as instance t, went through a tokenization procedure handled by different pre-processing functions depending on their language; we used BERTweet's module [12] for English and RoBERTuito's repository [13] for Spanish. Additionally, we replaced the string of characters "&" with "&".
We also included an additional feature regarding emotional information, in particular, we considered the categorical model of emotions [14], and applied BERT models [15, 16] fine-tuned with a Twitter Sentiment Analysis dataset [17] to assign the most likely prevailing emotional category to each instance t.

Used Features. After the pre-processing step, we conducted our experiments using the following features:

- **Text of the tweet (t)**: raw contents of the tweet.
- **Country**: The source country of the diplomat who posted the tweet.
- **Type**: The way how the tweet was disseminated: tweet, retweet, reply or quote.
- **Emotion**: Emotional category assigned according to the corresponding pre-trained language model, as described above.

BERT-based classifiers. Our approach relies on Bidirectional Encoder Representations from Transformers (BERT) models, which allowed us to create pre-trained language representations combining left and right contexts (thus generating a deep bidirectional Transformer) [18]. We used a BERT-based model pre-trained on tweets for each of the two languages:

- **BERTweet** is a widely available large-scale language model pre-trained on 850 million tweets in English [19]. The RoBERTa [20] pre-training process with a masked language modeling objective was used to train BERTweet.
- **RoBERTuito** is a pre-trained language model for content in Spanish [21], trained on 500 million tweets, also using the RoBERTa's pre-training process.

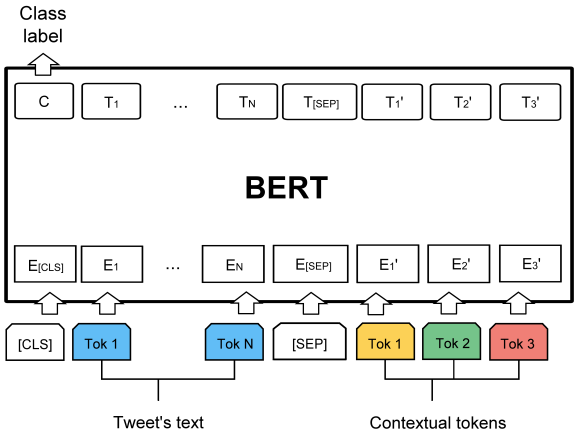

Figure 2: BERT's auxiliary input diagram with the contextual features concatenated to the tweet's text (adapted from [18]).

Figure 2 illustrates the process to combine a tweet t with its corresponding set of associated contextual features. We took advantage of BERT models' possibility of adding more tokens as an auxiliary input in a similar fashion as [22, 23].

2.2. Task 2: Multi-class propaganda characterization

Unlike Task 1, for addressing Task 2 we did not take advantage of any Transformer-based model or contextual information, instead, we used a more traditional method. Under the intuition that instances belonging to the same propaganda category are likely to contain similar words, we decided to exploit a *k-Nearest Neighbors (kNN)* classifier, together with a Boolean bag-of-words representation. For pre-processing purposes, we replaced URLs, hashtags, and mentions with a label "URL", "HASHTAG", and "MENTION", respectively. In addition, the punctuation marks were removed. Provided that this is a multi-label task, instead of assigning the majority-voting class determined by *kNN*, we associated all the categories of its neighboring instances. It is important to mention that, for an unlabeled instance to be tagged with a category (or a set of them), first it must be recognized as propagandist by our binary model; see Section 2.1. Figure 3 shows a schematic representation of the label assignment process for a given instance t according to the k-value chosen.

3. Experimental Setting

3.1. Data

The datasets are a collection of $1,292,885$ tweets published by 163 Chinese, 217 Russian, 283 European Union, and 314 United States authorities between January 1st, 2020, and March 11th, 2021 [24]. In both datasets, a further reduction of the sample was based on the strategic narratives' theory [25], sampling tweets that contained different terms associated with identity, system, and issue narratives. Table 1 shows the distributions of the train and test sets for both

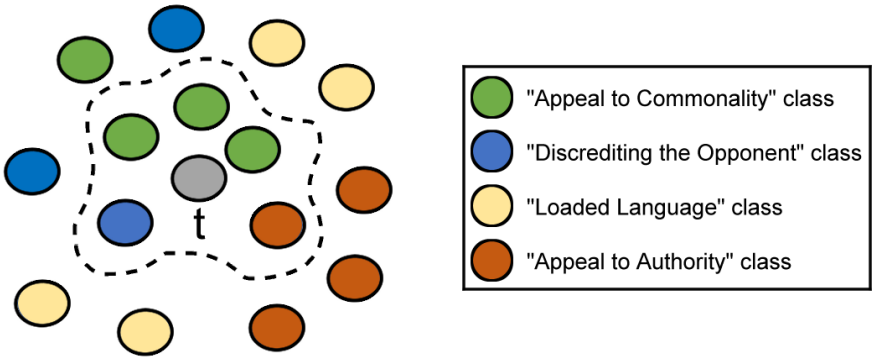

Figure 3: For a given propagandist instance t (in gray), we identify its 5 *nearest-neighbors*, and then we assign it the label(s) included in its neighborhood. In this case, t would be associated to all three categories: *"Appeal to Commonality"*, *"Discrediting the Opponent"*, and *"Appeal to Authority"* instead of only the most frequent one (i.e. *"Appeal to Commonality"*).

English and Spanish (refer to [3] for further information). The last section of the table reports the emotion distribution, inferred as described in Section 2.1.

3.2. Classifiers

Our models were implemented using Python 3.7 [26], Scikit-learn [27], and the HuggingFace library [28]. Table 2 reports the hyperparameters used for BERTweet and RoBERTuito. For Task 2, we used a customized implementation of the kNN classifier [29] with cosine as the similarity metric. The value of k was set to 5.

3.3. Runs' Configuration

As mentioned before, in our approach, we take advantage of the auxiliary input feature of the transformer models. For participating in the shared task, we proposed five different configurations:

- **Run 1** - Text alone.
- **Run 2** - Text + Country.
- **Run 3** - Text + Type.
- **Run 4** - Text + Emotion.
- **Run 5** - Text + a customized combination of contextual features.

For the latter, during development we assessed the performance of different combinations of the contextual features. The best-performing ones for each language were chosen: for English, *Country + Type*, while for Spanish *Country + Type + Emotion*.

Table 1
Data distribution for the English and Spanish corpora.

Class	Train set (ENG)	Test set (ENG)	Train set (SPA)	Test set (SPA)
Propaganda	1,974	N/A	1,199	N/A
Non-propaganda	6,434	N/A	4,921	N/A
Country				
China	2,170	852	2,178	819
European Union	2,043	873	1,508	957
Russia	2,005	955	795	596
USA	2,190	924	1,639	1,099
Type of tweet				
Tweet	6,742	2,856	3,586	2,302
Quoted	825	356	888	541
Retweet	473	227	1,221	401
Reply	368	165	425	227
Emotion*				
Anger	2,270	760	259	90
Fear	276	72	5	4
Joy	5,216	2,569	649	376
Love	114	53	N/A	N/A
Others	N/A	N/A	4,961	2,919
Sadness	508	141	224	66
Surprise	24	9	22	16
Total	8,408	3,604	6,120	3,471

* As inferred by our in-house model.

Table 2
Hyperparameter settings for both BERTweet and RoBERTuito.

Parameter	Value
Batch size	32
Learning rate	2e-5
Number of epochs	3
Max sequence length	250
Optimizer	Adam

4. Results

4.1. Development stage

We evaluated our models under a stratified 5-fold cross-validation scheme. Table 3 shows the results obtained for Task 1 during the development stage. It can be noted that adding contextual information allows for improving classification scores for all the experiments in comparison to Run 1, which uses the text alone.

Table 3
Obtained result in terms of F1-score during the development stage in both languages.

Run	Added features	macro F1 (ENG)	macro F1 (SPA)
Run 1	None	0.799	0.795
Run 2	Country	0.812	**0.812**
Run 3	Type	0.805	0.810
Run 4	Emotion	0.816	0.799
Run 5	Combination	**0.817**	0.810

4.2. Official Results

Table 4 shows the scores obtained by our official submissions. Our approach achieved very competitive results, in particular in Task 1. It ranks at the first and second positions for Spanish and English, respectively. It is worth noting that, in the complete list of results and position ranking [3], our **Run 1** was generally positioned below the rest of our contextualized runs.

The boxplots in Figure 4 show the distribution of the runs (in terms of F1-score) submitted by all the participant teams at the shared task. We also observe that due to a greater number of entries in English than in Spanish, there is a greater spread of data and a positive skew for this modality. Our strategy for Task 2 was positioned within the inter-quartile range below the median. We attribute this outcome to the simplicity of the proposed approach.

Table 4
Official results obtained by our best-performing runs in the shared task.

Task	Rank	Run	ICM-Hard[30]	ICM-Hard Norm[30]	F1
Task 1 SPA	1 of 18	Run 3	0.1724	0.8421	0.6681
Task 1 ENG	2 of 30	Run 4	0.1957	0.8180	0.6777
Task 1 AVG	1 of 16	Run 4	0.1576	0.8196	0.6501
Task 2 SPA	10 of 17	Run 3	-0.2344	0.8892	0.3761
Task 2 ENG	17 of 28	Run 5	-0.1239	0.9148	0.3866
Task 2 AVG	9 of 15	Run 3	-0.1487	0.9008	0.3824

4.3. Analysis

Attempting to figure out how adding specific contextual information impacts positively the classification performance, we decided to manually analyze a subset of instances. Some examples are shown in Table 5. In the first example, despite the multiple references to "Russia", and also that the text presents a nationalist tone, without the help of contextual information about the country of origin of the author of the tweet the classifier was not able to adequately recognize it as *propagandist*. In the second example, providing information about how the tweet was disseminated helps the classifier to recognize propaganda in the instance which did not happen when using only the raw text. In the third case, there is a message with a positive connotation in the ecological domain, which had not been properly identified as propaganda until emotional information was added. Finally, in the last example, the former president of the US criticizes the current president Joe Biden for some of his foreign policies; even when the propagandist

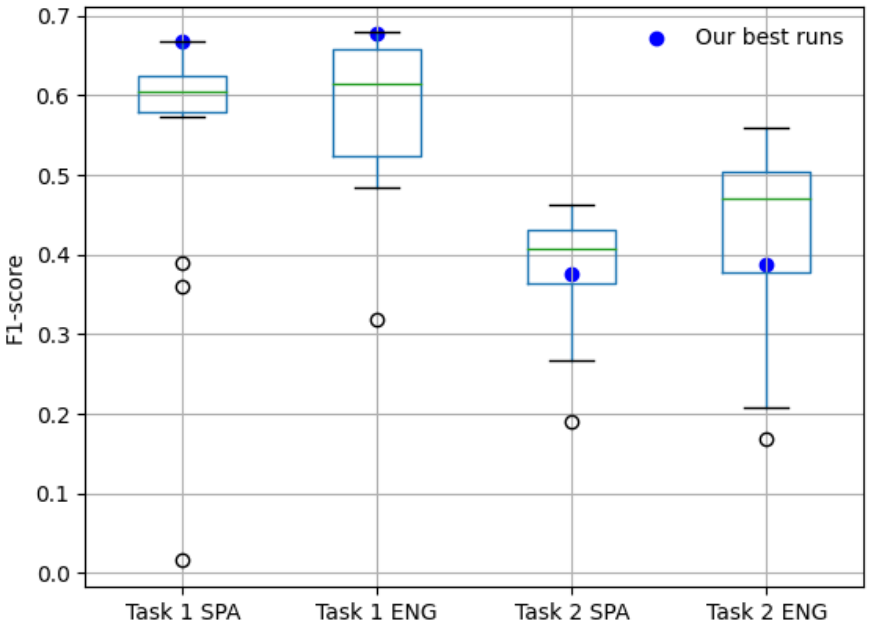

Figure 4: Box plots of the results for each task.

sense of this instance is more evident, the textual-based classifier was not able to identify it; however, this changes when information regarding the Type and Country of the post is added.

5. Conclusions

This paper describes our participation in the DIPROMATS shared task. The proposed approach showed an outstanding performance, ranking in the best positions in both English (2^{nd}) and Spanish (1^{st}). In particular, we focused on combining text messages with contextual information, under the intuition that propaganda emerges differently based on a variety of contextual elements. We attribute the success of our proposal on two main aspects: 1) The selection of relevant models that are particularly created for Twitter data, and 2) The usefulness of the features regarding contextual information during classification. From our results, we consider that in some way propaganda detection can be improved by providing contextual information to the classifiers to treat the problem more specifically. It seems that, complementing posts with their geographical origin, predominant emotion, and how they were disseminated, can be helpful to make decisions regarding the potential presence of propaganda. As future work, we are interested in further explore the use of other kinds of contextual features for detecting propagandist content, such as for example the political bias of the sources.

Table 5
Examples of tweets in training set where adding context features resulted in a right prediction. Words in bold represent those that may have had an impact when combined with contextual features.

Text	Class	Prediction w/o context	Prediction w/context	Type of context
(emoji of letter i) #Crimea is a vital part of (emoji of **russian** flag) **Russian** civilization. It is the point of origin of **Russian** Christianity (988), was in Ancient Rus (X-XII cent.) & **Russian** Empire (1783-1917), Soviet **Russia** & USSR (1917-91), re-united with #**Russia** (2014 - ∞) - dear to the hearts of all **Russians** URL	Prop	Non-prop	Prop	Country: Russia
@USER There is a Chinese saying *Under the over-turned nest, there are complete eggs** meaning if a bird's nest is destroyed, how is it possible to have safe eggs? If a country can not defend itself, people will suffer hugely. Both history and the present is proof.	Prop	Non-prop	Prop	Type: Reply
Today marks a **major milestone** in making Europe the first climate neutral continent in the world. With the new target to cut EU greenhouse gas emissions by at least 55% by 2030, we will lead the way to a **cleaner planet** and a **green recovery**.	Prop	Non-prop	Prop	Emotion: Joy
Joe Biden delivered remarks to union members after spending 47 years **giving their jobs to China and foreign countries** in exchange for **campaign** cash...	Prop	Non-prop	Prop	Type: Tweet, Country: USA

*The original saying was typed in Chinese, we offer a translation to English.

References

[1] C. Sparkes-Vian, Digital Propaganda: The Tyranny of Ignorance, Critical Sociology 45 (2019) 393–409. URL: https://doi.org/10.1177/0896920517754241. doi:10.1177/0896920517754241. arXiv:https://doi.org/10.1177/0896920517754241.

[2] DIPROMATS, Automatic Detection and Characterization of Propaganda Techniques from Diplomats homepage, https://sites.google.com/view/dipromats2023/home, 2023. [Online; accessed 30-May-2023].

[3] 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).

[4] G. Da San Martino, A. Barrón-Cedeño, P. Nakov, Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection, in: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 162–170. URL: https://aclanthology.org/D19-5024. doi:10.18653/v1/D19-5024.

[5] V. Vorakitphan, E. Cabrio, S. Villata, "Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 1498–1507. URL: https://aclanthology.org/2021.ranlp-1.168.

[6] O. Balalau, R. Horincar, From the Stage to the Audience: Propaganda on Reddit, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3540–3550. URL: https://aclanthology.org/2021.eacl-main.309. doi:10.18653/v1/2021.eacl-main.309.

[7] M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, Improving Cyberbullying Detection with User Context, in: P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, E. Yilmaz (Eds.), Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 693–696. doi:https://doi.org/10.1007/978-3-642-36973-5_62.

[8] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, Mean Birds: Detecting Aggression and Bullying on Twitter, in: Proceedings of the 2017 ACM on Web Science Conference, WebSci '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 13–22. URL: https://doi.org/10.1145/3091478.3091487. doi:10.1145/3091478.3091487.

[9] M. Ribeiro, P. Calais, Y. Santos, V. Almeida, W. Meira Jr., Characterizing and Detecting Hateful Users on Twitter, Proceedings of the International AAAI Conference on Web and Social Media 12 (2018). URL: https://ojs.aaai.org/index.php/ICWSM/article/view/1505. doi:10.1609/icwsm.v12i1.15057.

[10] M. Casavantes, M. E. Aragón, L. C. González, M. Montes-y Gómez, Leveraging posts' and authors' metadata to spot several forms of abusive comments in Twitter , Journal of Intelligent Information Systems (2023) 1–21.

[11] S. NP, B. Muniyal, D. Teja, L. Maben, Detection of Political Intent Through Analysis of Tweets and Homophily Elements (2020).

[12] BERTweet, TweetNormalizer.py, https://github.com/VinAIResearch/BERTweet/blob/master/TweetNormalizer.py, 2021. [Online; accessed 30-May-2023].

[13] RoBERTuito, A pre-trained language model for social media text in Spanish, https://huggingface.co/pysentimiento/robertuito-base-uncased, 2022. [Online; accessed 30-May-2023].

[14] P. Ekman, Universals and cultural differences in facial expressions of emotion., in: Nebraska symposium on motivation, University of Nebraska Press, 1971.

[15] Model Description,bert-base-uncased-emotion, https://huggingface.co/bhadresh-savani/

bert-base-uncased-emotion, 2018. [Online; accessed 30-May-2023].
[16] J. M. Pérez, J. C. Giudici, F. Luque, pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, 2021. arXiv:2106.09462.
[17] "Emotion", Dataset Summary, https://huggingface.co/datasets/philschmid/emotion, 2022. [Online; accessed 30-May-2023].
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).
[19] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
[20] 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, 2019. arXiv:1907.11692.
[21] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.
[22] J. A. Fuentes-Carbajal, M. Montes-y Gómez, L. Villaseñor-Pineda, Does This Tweet Report an Adverse Drug Reaction? An Enhanced BERT-Based Method to Identify Drugs Side Effects in Twitter, in: Pattern Recognition: 14th Mexican Conference, MCPR 2022, Ciudad Juárez, Mexico, June 22–25, 2022, Proceedings, Springer, 2022, pp. 235–244.
[23] F. Sánchez-Vega, A. P. López-Monroy, BERT's Auxiliary Sentence focused on Word's Information for Offensiveness Detection., 2021.
[24] DIPROMATS, Automatic Detection and Characterization of Propaganda Techniques from Diplomats data page, https://sites.google.com/view/dipromats2023/data, 2023. [Online; accessed 30-May-2023].
[25] B. O'Loughlin, A. Miskimmon, L. Roselle, Strategic Narratives: Communication Power and the New World Order, 2013. doi:10.4324/9781315871264.
[26] G. Van Rossum, F. L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[28] T. Wolf, 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, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.
[29] ML-algos, KNeighborsClassifier.py, https://github.com/turnerluke/ML-algos/blob/main/knn/KNeighborsClassifier.py, 2022. [Online; accessed 30-May-2023].
[30] 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.