PropaLTL at DIPROMATS 2024: Cross-lingual Data Augmentation 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 and Narratives from Diplomats of Major Powers* shared task (DIPROMATS). For this edition, we experimented with data augmentation, leveraging both English and Spanish training sets in a cross-lingual setting. As in the previous edition, the use of contextual features of the posts was also considered to improve their interpretation and subsequent classification. Our results show a slight increase in classification performance by incorporating more training instances. In particular, our strategy of cross-lingual data augmentation obtained competitive scores in the *binary propaganda identification task*: the eighth position for English out of 26 runs, and the eighth position for Spanish out of 30 runs.

Keywords

Propaganda detection, contextual information, data augmentation, transformers

1. Introduction

Propaganda can be defined as "an evolving set of techniques and mechanisms which facilitate the propagation of ideas and actions" [1]. The subtlety of propaganda enables it to function as a sophisticated method of manipulation, as its information does not have to be false, and its characteristics might only become evident after thorough observation, which sets propaganda apart from disinformation that can be debunked through fact-checking [2]. One of the objectives of DIPROMATS at IberLEF 2024 [3] is the identification of propaganda techniques; thus, a key focus lies in discerning hostile, deceptive, and emotionally charged claims [4]. One of the most distinctive aspects of this shared task lies in the composition of its corpus, which consists of Spanish and English tweets written by official government accounts, ambassadors, as well as other diplomatic profiles like consuls and missions from China, Russia, the United States, and the European Union.

IberLEF 2024, September 2024, Valladolid, 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)

Our strategy this year is focused on *binary propaganda identification*, which involves deciding whether a tweet contains propagandist content. In addition to resuming aspects of our participation in the 2023 edition (i.e., exploiting contextual information [5]), this year we have included a cross-lingual data augmentation configuration. Our intuition is that, due to the modest size of the datasets provided for training, models could benefit from more training data. In order to use data as close as possible to the task domain, we leveraged the tweets in both Spanish and English, translating them and using them crosswise. We have also implemented a filtering stage to select those tweets that would potentially benefit the cross-training process. Regarding the use of contextual features, as in the previous edition [2], we considered three aspects of a given tweet: (i) country of origin of its author, (ii) way in which it has been disseminated (i.e., original tweet, retweet, quote, or reply), and (iii) most likely emotion that it evokes (inferred with a pre-trained supervised model). We used BERT models' 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 conclusions about our participation in the shared task.

2. Proposed Approach

When designing our propaganda detection models, it was noted that a potential factor hindering the classifiers' performance could be data scarcity. In order to investigate this, we conducted an initial experiment using the data from DIPROMATS 2024 to assess how the classification performance was affected when the proportions of training data were altered. For this experiment, 7,146 tweets were considered for the English training set and 5,202 tweets for Spanish. These volumes represented training with "100%" data. Then, using stratified random sampling, smaller volumes of data were created: 25%, 50% and 75%. With the four aforementioned data volumes, BERT-based classifiers [6] were trained and predictions were made 5 different times (reporting their average) over a custom test set (see Section 4.1). The results of this experiment in Figure 1 show a consistent pattern of improvement in F1-score over the propaganda class as the volume of data increases in both languages. Upon analyzing these findings, we hypothesized that increasing the quantity of training data for each language could potentially enhance the classifiers. Accordingly, we developed our solution based on cross-lingual data augmentation, seeking to use more tweets belonging to the same domain (i.e., diplomat's propaganda), the decision was made to consider the data sourced from the DIPROMATS task pertaining to opposite languages.

2.1. Subtask 1A: Binary Propaganda Identification

Our approach to Subtask 1a consists of three main modules: data preparation, features computation, filtering process, and BERT-based classification.

Data preparation. We extracted the country of each tweet (which corresponds to that of the user who posted it) as well as the kind of post (how it was disseminated). The text of each tweet, from now on referred to as t, was passed through a tokenization procedure handled

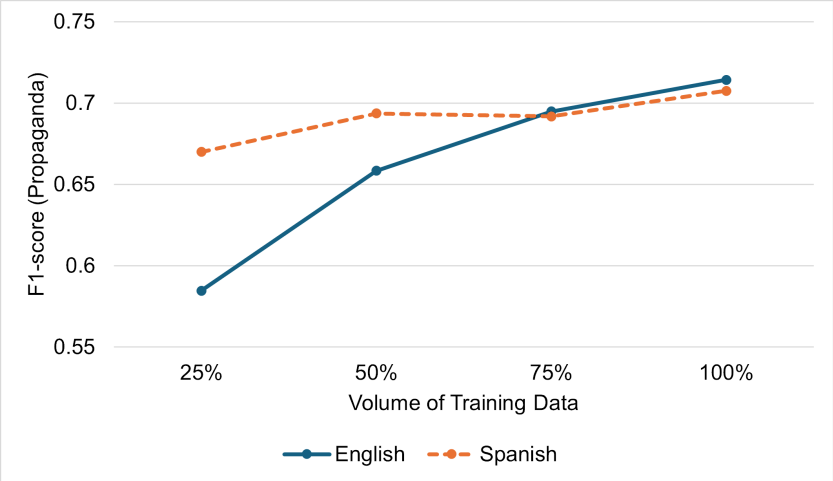

Figure 1: F1-scores over the propaganda class of classifiers while varying the volume of training data.

by different pre-processing functions depending on their corresponding language[1]. We also included an additional feature regarding emotional information, in particular, we considered the categorical model of emotions [9], and applied BERT models [10, 11] fine-tuned with a Twitter Sentiment Analysis dataset [12] to assign the most likely prevailing emotional category to each tweeet t. Before using the datasets for cross-lingual experiments, it was necessary to create translated versions of both. Based on the work by [13], OPUS-MT models [14] were used to perform machine translations[2].

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 person 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.

Filtering process. The methodology employed for augmenting data to the English train set is as follows. Initially, a series of five BERTweet models are trained using all available tweets in English. Subsequently, these models are used to generate predictions over the train set in Spanish (previously translated to English). If a minimum of three out of the five predictions align with the true label of the tweet (originally in Spanish), it is selected for inclusion in the

[1]We used BERTweet's tokenizer module [7] for English and RoBERTuito's repository [8] for Spanish.
[2]For Spanish to English we take advantage of https://huggingface.co/Helsinki-NLP/opus-mt-es-en. On the other hand, for English to Spanish https://huggingface.co/Helsinki-NLP/opus-mt-en-es.

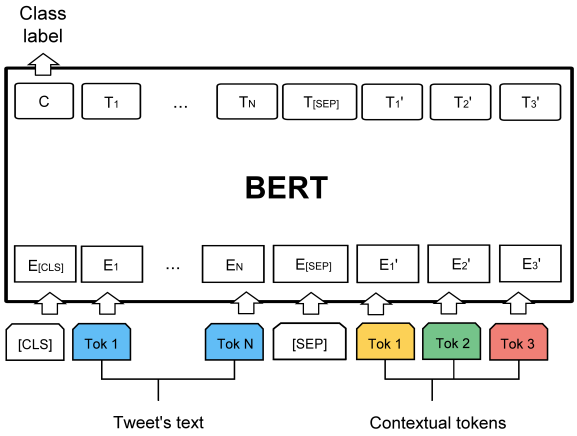

Figure 2: BERT's auxiliary input diagram with the contextual features concatenated to the tweet's text (adapted from [6]).

"augmented" English train set. The same procedure is applied in the opposite direction to augment the Spanish training set.

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) [6]. We used a BERT-based model pre-trained on tweets for each language:

- **BERTweet** is a widely available large-scale language model pre-trained on 850 million tweets in English [15]. The RoBERTa [16] pre-training process with a masked language modeling objective was used to train it.
- **RoBERTuito** is a pre-trained language model for content in Spanish [17], trained on 500 million tweets, also using the RoBERTa pre-training process.

Figure 2 illustrates the combination of a tweet t with its corresponding associated contextual features. We took advantage of BERT models' possibility of adding more tokens as an auxiliary input similarly as [18, 19].

3. Experimental Setting

3.1. Data

The datasets are a collection of tweets published by authorities from China, Russia, the European Union, and the United States between January 1st, 2020, and March 11th, 2021 [20]. Table 1 shows the distributions of the train and test sets for both English and Spanish (refer to [21] for further information). The last section of the table reports the emotion distribution, inferred as described in Section 2.1.

Table 1
Data distribution for the English and Spanish corpora. For the distribution of emotional categories, "Love" was exclusive to English while "Others" was exclusive to Spanish.

Class	Train set (en)	Test set (en)	Train set (es)	Test set (es)
Propaganda	1,971	N/A	1,196	N/A
Non-propaganda	6,437	N/A	4,924	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 models.

3.2. Classifiers

We used a BERTweet based classifier for English and a RoBERTuito for Spanish. These models were implemented using Python 3.7 [22], and the HuggingFace library [23]. The hyperparameters used for both classifiers were a batch size of 32, learning rate of 2e-5, 3 epochs, max sequence length of 250, and Adam optimizer.

3.3. Runs' Configuration

As mentioned before, in our approach, we took advantage of both translated propaganda instances and the auxiliary input of the transformer models. For our participation in the shared task, we submitted the following five different configurations:

- **Run 1** - Vanilla. No data augmentation, no contextual attributes.
- **Run 2** - With contextual attributes. Country + type + emotion.
- **Run 3** - Data augmentation. Adding all translated propaganda instances from the other language.

Table 2
Obtained results in terms of F1-score (propaganda class) during the development stage in both languages, using our custom train and test sets.

Run	Added features	F1-True (en)	F1-True (es)
Run 1	None	0.7142 ± 0.024	0.7074 ± 0.044
Run 2	Context	$\mathbf{0.7246 \pm 0.011}$	$\mathbf{0.7453 \pm 0.009}$
Run 3	All translated propaganda	0.6965 ± 0.028	0.6624 ± 0.044
Run 4	Filtered translated propaganda	0.7227 ± 0.008	0.7131 ± 0.020
Run 5	Filtered translated propaganda + Context	0.6864 ± 0.024	0.7183 ± 0.035

- **Run 4** - Data augmentation. Adding filtered translated propaganda instances, as described in Section 2.1. The number of tweets from the Spanish train set that were added to the English train set was 498, while the number of tweets coming from the train set in English that were added to the train set in Spanish was 1030.
- **Run 5** - Data augmentation. A combination of Run 4 and Run 2.

4. Results

4.1. Development stage

During the development phase, our approach involved splitting the training sets (for both English and Spanish) into two fixed partitions: 85% allocated to a training partition (which we will subsequently refer to as "custom train set"), and the remaining 15% designated for testing (hereafter referred to as "custom test set"). Table 2 shows the results for Task 1, which correspond to the average of 5 executions of BERTweet/RoBERTuito along with their respective standard deviations.

4.2. Official Results

Table 3 shows the scores obtained by our official submissions. Our approach achieved competitive results in Subtask 1a. It ranks at the eighth position for both English and Spanish. It is worth noting that, in the complete list of results and position ranking [21], our Run 5 was generally positioned below the rest of our runs. This leads us to think that, using our methods, data augmentation combined with the use of all contextual attributes are not complementary strategies. In light of the results showing superior performance by Run 4 over Run 2, contrary to our expectations from Table 2, it is possible that the addition of contextual features may have led to overfitting in the models.

The boxplots in Figure 3 show the distribution of the runs, in terms of F1-score over the propaganda class, submitted by all the participant teams at the shared task. Our cross-lingual augmented runs are positioned either slightly above or close to the upper quartile, while our vanilla runs are between the median and upper quartile.

Table 3
Official results obtained by our best-performing runs in the shared task.

Task	Rank	Run	ICM[24]	F1-True
Task 1 es	8 of 26	Run 4	0.1802	0.6718
Task 1 en	8 of 30	Run 4	0.1493	0.6455
Task 1 AVG	6 of 33	Run 4	0.1667	0.6614

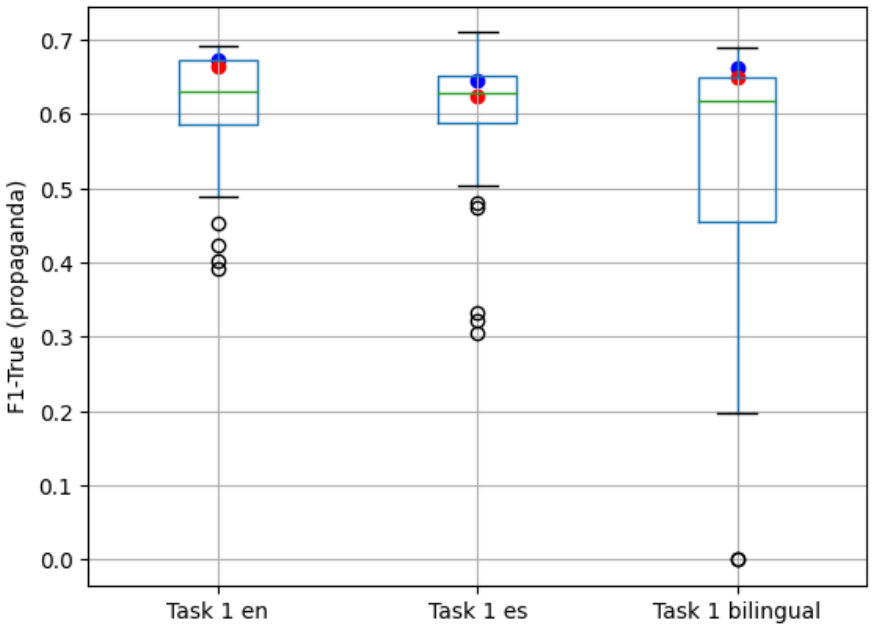

Figure 3: Box plots of the results for each task. The blue dots represent our best runs (filtered data augmentation), while the red dots represent our vanilla runs (no data augmentation nor added context).

Attempting to explain the rationale behind the modest performance differences observed in our models when using or not data augmentation, we conduct supplementary analyses. The first one is based on the speculation that perhaps both English and Spanish train sets shared Twitter accounts, and subsequently some tweets used to augment the training sets are likely to both come from these common accounts and contain similar propaganda. However, our examination refuted this first hypothesis at least for the data augmentation in Spanish, showing that there is no strong overlap of accounts in both languages. In summary we observed the following:

- It turned out that the number of unique Twitter accounts in the English Train set was 491, while the number in the Spanish Train set was 128.
- The total number of accounts in common (in other words, the intersection of these sets) is 39.
- Of the 498 Spanish tweets that were added to the English Train set, the number of tweets

belonging to the common accounts was 284 (more than half).

- Of the 1030 English tweets that were added to the Train set in Spanish, the number of tweets that belong to the common accounts was 82.

Additionally, we calculated Kullback–Leibler divergences [25, 26] between term frequencies of original and augmented sets for each language. For English, the divergence between the custom train set and the augmented train set was 0.0599, while for Spanish the divergence was 0.1636. The divergence in Spanish is greater than the divergence in English, possibly because twice as much data was augmented in Spanish as in English; however both values are rather small. Both results, which are very close to 0, indicate that the vocabulary frequency distributions between the sets had very few differences and, therefore, that the collections in both languages have similar tweets, possibly corresponding to common propaganda campaigns.

5. Conclusions

This paper describes our participation in the 2024 DIPROMATS shared task. The proposed approach showed good performance in both English (8^{th} position) and Spanish (8^{th} position). In particular, we focused on cross-lingual data augmentation, while also resuming part of our previous participation by combining text messages with contextual information. Through our participation, it has been confirmed that improvements in the classification outcomes can be achieved for this subtask, by incorporating additional training resources, particularly sourced from the same domain (diplomatic entities) and time period. As future work, we would like to explore the idea of continuing to incorporate different data collections and determine if other types of propaganda are "compatible" with those in the DIPROMATS corpus.

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] 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] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[4] DIPROMATS, Automatic Detection and Characterization of Propaganda Techniques and Narratives from Diplomats of Major Powers - homepage, https://sites.google.com/view/dipromats2024/home, 2024. [Online; accessed 06-June-2023].

[5] M. Casavantes, M. Montes-y-Gómez, D. I. H. Farías, L. C. González-Gurrola, A. Barrón-Cedeño, PropaLTL at DIPROMATS: Incorporating Contextual Features with BERT's Auxiliary Input for Propaganda Detection on Tweets, in: M. Montes-y-Gómez, F. Rangel, S. M. J. Zafra, M. Casavantes, B. Altuna, M. Á. Á. Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. J. Escalante, M. Á. G. Cumbreras, J. A. García-Díaz, J. Á. G. Barba, R. L. Tamayo, S. Lima, P. Moral, F. M. P. del Arco, R. Valencia-García (Eds.), *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, Jaén, Spain, 2023. URL: https://ceur-ws.org/Vol-3496/dipromats-paper2.pdf.

[6] 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).

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

[8] 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].

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

[10] Model Description,bert-base-uncased-emotion, https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion, 2018. [Online; accessed 30-May-2023].

[11] J. M. Pérez, J. C. Giudici, F. Luque, pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, 2021. arXiv:2106.09462.

[12] "Emotion", Dataset Summary, https://huggingface.co/datasets/philschmid/emotion, 2022. [Online; accessed 30-May-2023].

[13] V. Ahuir, L. F. Hurtado, F. García-Granada, E. Sanchis, ELiRF-VRAIN at DIPROMATS 2023: Cross-lingual Data Augmentation for Propaganda Detection, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, Jaén, Spain, 2023. URL: https://ceur-ws.org/Vol-3496/dipromats-paper6.pdf.

[14] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, M. L. Forcada (Eds.), Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] F. Sánchez-Vega, A. P. López-Monroy, BERT's Auxiliary Sentence focused on Word's Information for Offensiveness Detection., 2021.

[20] DIPROMATS 2024, Task 1 Data page, https://sites.google.com/view/dipromats2024/task-1/data, 2024. [Online; accessed 06-June-2023].

[21] P. Moral, J. Fraile, G. Marco, A. Peñas, J. Gonzalo, Overview of DIPROMATS 2024: Detection, Characterization and Tracking of Propaganda in Messages from Diplomats and Authorities of World Powers, Procesamiento del Lenguaje Natural 73 (2024).

[22] G. Van Rossum, F. L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[23] 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.

[24] 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.

[25] S. Kullback, R. A. Leibler, On Information and Sufficiency, The Annals of Mathematical Statistics 22 (1951) 79 – 86. URL: https://doi.org/10.1214/aoms/1177729694. doi:10.1214/aoms/1177729694.

[26] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.