ELiRF-VRAIN at DIPROMATS 2023: Cross-lingual Data Augmentation for Propaganda Detection

Vicent Ahuir, Lluís Felip Hurtado, Fernando García-Granada* and Emilio Sanchis

Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia. Spain

Abstract

In this paper, we present our approach to the IberLEF 2023 DIPROMATS shared task. Our system is based on Deep Neural Network models that use pre-trained classification BERT-like models, that have been fine-tuned by using the corpus supplied by the organization and a data augmentation process. We have developed one main model for each subtask and language (Spanish and English). Based on these initial models, a process to correct inconsistencies in the labeling has been proposed and the use of cross-lingual models has also been tested. Results confirm that the proposed approach is adequate for the problem.

Keywords

Transformers, Propaganda Detection, Cross-lingual, Data Augmentation

1. Introduction

Nowadays, there are many ways to spread information on the internet to influence the thought or opinions of people. Fake news can be used for this purpose. However, more sophisticated methods, such as propaganda, are complicated to discriminate from factual information. In fake news, objective information can be used to check the veracity, even by considering the source of information. However, a thin line separates propaganda from real information or opinion.

Propaganda aims to influence people by focusing the message on specific aspects that can be true but are not the real description of the notice. There are many techniques developed to generate successful propaganda messages. Some techniques, such as slogans, repetitions, or casual oversimplification, are described in [1]. There are some previous attempts to develop corpus and systems for detecting propaganda in texts, such as [2, 3, 1].

The propaganda classification problem can be addressed from two points of view: a binary classification of the text in terms of propaganda/non-propaganda and detection of the techniques used in the propaganda texts. Both tasks have been proposed in this competition. Our proposed

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

viahes@dsic.upv.es (V. Ahuir); lhurtado@dsic.upv.es (L. F. Hurtado); fgarcia@dsic.upv.es (F. García-Granada); esanchis@dsic.upv.es (E. Sanchis)

https://vrain.upv.es/elirf/ (V. Ahuir); https://vrain.upv.es/elirf/ (L. F. Hurtado); https://vrain.upv.es/elirf/ (F. García-Granada); https://vrain.upv.es/elirf/ (E. Sanchis)

0000-0001-5636-651X (V. Ahuir); 0000-0002-1877-0455 (L. F. Hurtado); 0000-0003-2213-4213 (F. García-Granada); 0000-0002-6737-4723 (E. Sanchis)

approach is based on Transformers [4] models that use pre-trained classification BERT-like [5] models that have been fine-tuned using the corpus supplied by the organization. We have developed two different classifiers, one for the binary classification problem of subtask-1 and the other for the multiclass classification problem of subtask-2. We present results for different systems, where the differences rely on the fine-tuned models. Considering the obtained results, we can conclude that the proposed approach is adequate to address the problem.

2. Dataset

The proposed challenge [6] consists of classifying tweets according to these two tasks:

- Task 1 - propaganda identification. This first subtask is a binary classification problem. The systems should decide whether a given tweet contains propaganda techniques.
- Task 2 - propaganda grain characterization. The second subtask aims to categorize the type of propaganda. It is a multiclass task where systems have to decide which available categories it fits, given a tweet. Four classes of propaganda (plus a negative) have been defined for this task (Not propagandistic, Appeal to Commonality, Discrediting the Opponent, Loaded Language, and Appeal to Authorit).
- Task 3 - fine-grained categorization. The third task aims to do a fine-grained categorization. It is a multiclass, multilabel task where systems have to decide which available categories it fits, given a tweet. 15 subclasses (plus a negative class) have been defined for this task: Flag Waving, Ad Populum / Ad antiquitatem, Name Calling, Undiplomatic Assertiveness / Whataboutism, Scapegoating, Propaganda Slinging, Appeal to Fear, Demonization, Personal Attacks, Doubt, Reductio Ad Hitlerum, Loaded Language, Appeal to False Authority and Bandwagoning.

The corpus provided by the organization consists of 6119 tweets in Spanish and 8408 tweets in English for the training set and 3471 tweets in Spanish, and 3604 tweets in English for the test set.

In addition to the text of the tweet, additional information is provided for each tweet, like the country of the diplomat that posted the tweet, the diplomat's username on Twitter, the tweet type (tweet, retweet, reply, or quote), the time when the tweet was posted (UTC), the sum of retweets and favorites that the tweet obtained and the main language employed in the tweet.

3. System description

In this work, we wanted to evaluate the capabilities of Transformers-based systems for the propaganda detection task. The decision of which pre-trained model to use was made through a previous validation process in which we compare the performance of several models pre-trained specifically for Spanish, English and other multilanguages available at the HuggingFace [7] public hub. Finally, the models chosen were the well-known BETO [8] for Spanish (https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased) and the updated monolingual version of TimeLM [9] for English (https://huggingface.co/cardiffnlp/twitter-roberta-base-sep2022).

We used data augmentation to increase the number of samples by translating Spanish samples into English and English ones into Spanish. We chose OPUS-MT models [10] for the translation process. Specifically, we use the following models available at the HuggingFace public hub: *Spanish-English* (https://huggingface.co/Helsinki-NLP/opus-mt-es-en), and *English-Spanish* (https://huggingface.co/Helsinki-NLP/opus-mt-en-es).

4. Fine-tuning process

To carry out the fine-tuning process of the pre-trained models, a random split of the training corpora of each language was performed. The validation partition for each language was made up of 20% of the corpus samples for that language. The training partition included the remaining 80% and the entire corpus of the other language translated as described in the previous section. Table 1 shows the results for the best epoch of each model on the validation set.

Table 1
Macro-averaging results on the validation set.

	Spanish (BETO)			English (TWITTER-EN)		
	P	R	F1	P	R	F1
Task1	86.97	85.29	86.09	81.29	83.60	82.33
Task2	62.99	59.62	61.24	87.44	86.10	86.60
Task3	65.39	0.47.83	53.09	62.83	65.23	63.65

The use of different models for the three tasks can produce labeling inconsistencies. For instance, for the same sample, the Task1 classifier can label a sample as not propaganda while the Task2 classifier can assigns it a propaganda label. To avoid this, we create an Discrepancy Correction Procedure (DCP). This procedure follows four steps:

1. Prioritize Task 1. If the Task1 classifier classifies a sample as *not propaganda*, the possibles labels for Task 2 and Task3 are deleted.
2. If there is not prediction for label Task 2 or Task 3, the sample is labeled as *not propaganda*.
3. If there is not common label for Task 2 and Task 3, the sample is labeled as *not propaganda*.
4. The labels for Task 2 and Task 3 are only the labels predicted by the models that are consistent with each other.

Figure 1 shows the *python* code for the Discrepancy Correction Procedure.

Furthermore, to test the models in a cross-lingual environment and have more variability, we have used the same translation models used for data augmentation to translate the test corpus of each language and label it with the model of the other language. Finally, we presented four runs to the shaded task in both languages:

- Run1: BETO for the Spanish test set and TimeML for the English test set. Each model finetuned individually for each task.
- Run2: Result of applying the DCP to the output of *Run1*.

Figure 1: Discrepancy Correction Procedure (DCP).

```
if sample.task1 is False:
    sample.task2 = None
    sample.task3 = None
else:
    if sample.task2 is None or sample.task3 is None:
        sample.task1 = False
        sample.task2 = None
        sample.task3 = None
    elif common_labels(sample.task2, sample.task3) is None:
        sample.task1 = False
        sample.task2 = None
        sample.task3 = None
    else:
        sample.task2 = common_labels_task2(sample.task2, sample.task3)
        sample.task3 = common_labels_task3(sample.task2, sample.task3)
```

- Run3: TimeML for the translated Spanish test set and BETO for the translated English test set.
- Run4: Result of applying the DCP to the output of *Run3*.

5. Results

Table 2 shows the results, in terms of Macro F1, obtained by our four runs in the challenge. To create a wider view of the results, it shows the highest score obtained by any team for each task and language. *Task1*, *Task2*, and *Task3* are the scores obtained by the systems when samples from both languages are taken into account: Spanish and English. Meanwhile, *_ES* (Spanish) and *_EN* (English) columns are the scores obtained on the specified task when only samples of the specified language are considered.

Table 2
Macro F1 of the four runs by task and language. *Highest* contains the highest values achieved in the competition by any team. Values in parentheses indicate the team's global position in the challenge and which run achieved the position.

	Task1	Task1_ES	Task1_EN	Task2	Task2_ES	Task2_EN	Task3	Task3_ES	Task3_EN
Run1	77.15	77.25	76.90	**49.29 (1)**	45.94	50.39	**36.34 (1)**	**39.43 (1)**	37.68
Run2	76.66	75.82	77.09 (6)	49.02	45.27	50.58 (3)	36.16	38.84	37.68
Run3	77.32 (5)	78.15 (5)	76.56	48.36	**46.26 (1)**	48.41	35.19	36.28	40.33 (3)
Run4	77.06	77.72	76.46	48.38	45.78	48.67	35.08	36.28	40.09
Highest	79.53	80.89	80.90	49.29	46.26	55.91	36.34	39.43	43.38

Results show that most of the best scores on multiclass or multiclass-multilabel tasks were achieved by *Run1*, which is the system that contains a model for each language, and no DCP is applied. With this run, we reach first place on *Task2* and *Task3* of the challenge. These results

indicate that adding the country to the input and the finetuning process with data augmentation benefited the good results. In the case of *Task1*, our best run achieved fifth place, where the system obtained only a 2.5% lower performance than the best results of the challenge, giving an idea of how tight the results were in the first task. In *Task1*, *Run1* did not get our best result; the *Run3* achieved this. Surprisingly, the system with cross-lingual classification got slightly better performance than *Run1*.

Analyzing scores by language, we observe that *Run3* tends to obtain the best scores in Spanish. This could indicate that the English model and (or) translations from Spanish to English get a better grammar structure and vocabulary that benefit these classification tasks. In the case of English, *Run2* is the one that obtains the best results in two of the three tasks (Task1 and Task2). Finally, results show that DCP slightly increased performance on some English classification systems (*Run1*, *Run2*) but decreased most of the other scores. Therefore, DCP did not have the desired effect, and further polishing is needed.

6. Conclusions

We have presented the ELiRF-VRAIN system for the DIPROMATS 2023 shaded task. Our approach is based in Transformers, and uses pre-trained BERT-like models. A characteristic of our system is that we used data augmentation taking advantage of the labeled samples of different languages. Additionally, a process to correct inconsistencies in the labeling and the use of cross-lingual models has also been tested. Results confirm that the proposed approach is adequate for the problem.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union" and “NextGenerationEU/MRR”, and by “ERDF A way of making Europe” under grants PDC2021-120846-C44 and PID2021-126061OB-C41. It is also partially supported by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training.

References

[1] G. D. S. Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news articles, CoRR abs/1910.02517 (2019). URL: http://arxiv.org/abs/1910.02517. arXiv:1910.02517.

[2] 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. URL: https://www.sciencedirect.com/science/article/pii/S0306457318306058. doi:https://doi.org/10.1016/j.ipm.2019.03.005.

[3] S. Yu, G. Martino, M. Mohtarami, J. Glass, P. Nakov, Interpretable propaganda detection in news articles, 2021, pp. 1597–1605. doi:10.26615/978-954-452-072-4_179.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach,

R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[6] 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).

[7] 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. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. 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://www.aclweb.org/anthology/2020.emnlp-demos.6.

[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[9] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251–260. URL: https://aclanthology.org/2022.acl-demo.25. doi:10.18653/v1/2022.acl-demo.25.

[10] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: 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.