CiTIUS at FakeDeS 2021: A Hybrid Strategy for Fake News Detection

Pablo Gamallo[1][0000−0002−5819−2469]

Centro de Investigación en Tecnoloxías da Información (CiTIUS) Universidade de Santiago de Compostela, pablo.gamallo@usc.gal

Abstract. This article describes several BERT-based supervised classification strategies submitted to Fake News Detection in Spanish Shared Task, where the sources of data are news annotated as fake or real. In our experiments, the systems were trained exclusively with the official datasets provided by the organizers of the shared task, without making use of any other source of information. The best system turned out to be a hybrid strategy that combines sentence similarity with some linguistic heuristics.

Keywords: Fake News · Transformers · BERT Sentence Similarity

1 Introduction

The widespread use of social media and e-communication platforms pushed people to rely on them as the main source for information. Unfortunately, an abnormal amount of fake news, rumours and disinformation has overflowed social media, with the aim of drawing the attention of their users to shape their opinions and judgments [10]. Fake news and all kind of disinformation can have dramatic effects on countries, businesses, and people on various levels, whether political or economically [2].

Many approaches have been proposed to identify the authenticity of published news on social media and e-communication platforms. Some of these approaches rely on the users of the platforms. For example, Facebook urges their users to report suspicious news or comments [14], and even makes uses of professionals to manually checks the reported comments and news published on their platform. The manual fact checking process also has been used by many other fact-checkers, journals and organizations to discover questionable news, however, this manual method is a waste of human efforts because of the huge amount of news published every second on social media [7]. Accordingly, automating the detection of fake news has caught the attention of researchers in academia and industry particularly after the incident of the American elections in 2016.

IberLEF 2021, September 2021, Málaga, Spain.

The use of Natural Language Processing (NLP) techniques with machine learning and deep learning methods for the detection of fake news can help to stop or at least reducing the misinformation. The use of both traditional machine learning techniques and deep learning approaches for the detection of hostile communication (a term that embraces both disinformation and hate speech) has received much attention recently [10, 1, 8], even though the most successful systems for such a task are those using domain-specific fine-tuning of pre-trained masked language models (i.e., Transformers architectures). For example, in the shared subtask focused on COVID-19 Fake News Detection in English, the best performances were achieved by Transformer-based systems [12].

In this paper, we describe the experiments performed to participate at Fake News Detection in Spanish Shared Task (FakeDeS 2021) [9] In order to train and develop the systems, the organizers of this task provide the Spanish Fake News Corpus, which consists of news compiled from different web sources and covering several topics, e.g. Economy, Science, Politics, Sport, Health, etc. [13]. The final test corpus provided by the organizers for evaluation contains news on COVID-19, a specific topic which was not including in the train and development corpora. The task consists of deciding if a piece of news is real or fake by considering both the title and body text of the news.

In this paper, in order to accomplish the Fake News Detection in Spanish Shared Task, we compare several strategies that make use of BERT-based models [4]. One of the proposed strategies is hybrid as they compute semantic similarity by combining BERT-based models with linguistic heuristics. This turned out to be our best model in the test dataset, and the sixth best model in the shared task, out of 21 participants.

The remaining of the paper is organized as follows. In section 2, we describe the proposed methods. Experimental results and discussion are addressed in section 3, while section 4 reports the conclusion of the current study as well as future work.

2 The Strategies

As it has been said, our aim is to compare several BERT-based strategies by making use of the training data provided by the organizers of Fake News Detection in Spanish Shared Task, which is part of Iberian Languages Evaluation Forum (IberLEF 2021) [11].

Bidirectional Encoder Representations from Transformers, known as BERT [4], is a bi-directional transformer-based language model learning information from left to right and from right to left. As any language model, it can be used to extract high quality language features from input text, but it can also be fine-tuned on specific NLP tasks such as entity recognition, classification, question answering, sentiment analysis, or claim verification in fact checking. In the experiments described in the next section, we will use BERT with three different strategies:

Fine-tune model: In the fine tuning strategy, we add a dense layer on top of the last layer of the pre-trained BERT model and then train the whole model by making use of the task specific dataset. Pre-trained BERT model has been fine-tuned in order to conduct the binary classification task to fake news detection.

Sentence similarity: BERT is also able to extract both contextualized word embeddings and sentence embeddings from text in order to compute semantic similarity between complex expressions or sentences. For the particular task at stake, we compare the target news of the test set with news labeled as *fake* in the training dataset and those labeled as *real*. Comparison consists in compute sentence similarity between the target news and all labeled news of the training data. The target news is classified as *fake* if the average of sentence similarities with fake news is exceeding the average of sentence similarities with real news. It is classified as *real* if the opposite situation arises.

Hybrid strategy: This uses the output of the previous method in order to identify the borderline cases, that is, those news that were classified as either fake or real with a low value (this value was set empirically). These borderline cases were then reclassified by using linguistic cues, such as size of the news, the presence of sentences written in capital letters, or specific fake statements (e.g., 5G...) in the body text of the news.

To compare these systems with traditional machine learning methods, we also performed tests with a Naive Bayes classifier. For this purpose, we used the system we implemented in a previous work for the task of bot detection [6], which relies on Linguakit [5] for tokenization and extraction of n-grams.

3 Experiments

3.1 Datasets

The datasets provided by the organizers consists of three partitions: train, development, and test. The train dataset consists of 676 labeled news containing 264K words. The development dataset consists of 295 news with 126K words, while the test dataset consists of 572 news with 190K words, the first being used for preliminary experiments and configuration of the systems, and the second one for final evaluation. The systems submitted to the shared task for final evaluation were trained by merging both the train and development datasets: 971 news with 391K words.

It should be noticed that we did not made use of any external source of knowledge or other annotated datasets to fine tune the models

3.2 Systems Configuration

Both fine tuning and sentence similarity strategies were performed with *BETO* [3], a pre-trained model for Spanish with 12 layers, by making use of Huggingface Transformers library [15]. Concerning fine tuning, the training dataset was

Methods	F1 devel	F1 test
BERT fine tuning	**76.55**	47.19
Sentence BERT	71.38	-
Sentence BERT + Heur.	73.54	**70.1**
Naive Bayes	68.58	-

Table 1. Results in F1 of four systems on both development and test datasets.

divided in both train (80%) and validation (20%) file partitions, while the development dataset was used as test file. For the official results, train and development were merged to build a larger training file.Concerning the use of BERT for sentence similarity, we built sentence embeddings with a pooling method: it adds a pooling operation to the output of the Transformer to derive fixed sized sentence embeddings. The specific pooling strategy we used is the mean of all output vectors.

3.3 Evaluation

Table 1 shows F1 scores (just for fake values) obtained by our four strategies on both development and final test datasets. The two strategies submitted to the shared task, which where the best in the development dataset, are in bold. The last column (F1 test) stands for the official scores obtained in the final evaluation. Our best system, the hybrid method, ranked 6th out of 21 participants. This system has a more stable behaviour than the fine tuned BERT model, whose F1 scores is much lower in the test dataset than in the development one. Traditional Naive Bayes classifier gives the lowest values on the development dataset. It is worth noting that the content of the test dataset is very different from that of the development one as they were built using different topics and journals of different countries, which makes test classification a very difficult task.

4 Conclusions

In this paper, several classification strategies have been compared for the fake news detection task on Spanish news. More precisely, we compared recent strategies relying on the use of Transformers, which provide deep semantic models with contextualized word embeddings. The best result combines BERT-based sentence similarity with linguistic heuristics. The results were submitted to Fake News Detection in Spanish Shared Task. Experiments were performed without considering external sources of knowledge or other annotated datasets.

In future work, we will try to carry out an in-depth analysis of the results obtained to establish what factors determine the significant difference between the different strategies with the datasets evaluated. We will also look for other sources of information and explore more linguistic information so as to improve the hybrid strategy.

References

1. Almatarneh, S., Gamallo, P., Pena, F.J.R., Alexeev, A.: Supervised classifiers to identify hate speech on english and spanish tweets. In: International Conference on Asian Digital Libraries. pp. 23–30. Springer (2019)
2. Apuke, O.D., Omar, B.: Fake news and covid-19: modelling the predictors of fake news sharing among social media users. Telematics and Informatics p. 101475 (2020)
3. Caete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Prez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
5. Gamallo, P., Garcia, M., Pieiro, C., Martinez-Castao, R., Pichel, J.C.: LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 239–244 (2018). https://doi.org/10.1109/SNAMS.2018.8554689
6. Gamallo, P., Almatarneh, S.: Naive-bayesian classification for bot detection in twitter. In: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2380/paper_194.pdf
7. Giachanou, A., Rosso, P.: The battle against online harmful information: The cases of fake news and hate speech. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 3503–3504 (2020)
8. Gilda, S.: Evaluating machine learning algorithms for fake news detection. In: 2017 IEEE 15th Student Conference on Research and Development (SCOReD). pp. 110–115. IEEE (2017)
9. Gómez-Adorno, H., Posadas-Durán, J.P., Bel-Enguix, G., Porto, C.: Overview of fakedes task at iberlef 2020: Fake news detection in spanish. Procesamiento del Lenguaje Natural **67**(0) (2021)
10. Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., Akbar, M.: Fake news detection using deep learning models: A novel approach. Transactions on Emerging Telecommunications Technologies **31**(2), e3767 (2020)
11. Montes, M., Rosso, P., Gonzalo, J., Aragn, E., Agerri, R., ngel lvarez Carmona, M., lvarez Mellado, E., de Albornoz, J.C., Chiruzzo, L., Freitas, L., Adorno, H.G., Gutirrez, Y., Zafra, S.M.J., Lima, S., de Arco, F.M.P., Taul, M.: Proceedings of the iberian languages evaluation forum (iberlef 2021). In: CEUR Workshop Proceedings (2021)
12. Patwa, P., Bhardwaj, M., Guptha, V., Kumari, G., Sharma, S., PYKL, S., Das, A., Ekbal, A., Akhtar, M.S., Chakraborty, T.: Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In: Chakraborty, T., Shu, K., Bernard, H.R., Liu, H., Akhtar, M.S. (eds.) Combating Online Hostile Posts in Regional Languages during Emergency Situation. pp. 42–53. Springer International Publishing, Cham (2021)

13. Posadas-Durán, J.P., Gómez-Adorno, H., Sidorov, G., Escobar, J.J.M.: Detection of fake news in a new corpus for the spanish language. Journal of Intelligent & Fuzzy Systems **36**(5), 4869–4876 (2019)
14. Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G.S., On, B.W.: Fake news stance detection using deep learning architecture (cnn-lstm). IEEE Access **8**, 156695–156706 (2020)
15. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, `https://www.aclweb.org/anthology/2020.emnlp-demos.6`