GDUF_DM at FakeDeS 2021: Spanish Fake News Detection with BERT and Sample Memory

Xixuan Huang[1], Jieying Xiong[1], and Shengyi Jiang[1,2(✉)]

[1] School of Information Science and Technology, Guangdong University of Foreign Studies, China
[2] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou
jiangshengyi@163.com

Abstract. Fake news widely spread on the Internet has had a negative impact on society. This article reports the solution of Spanish fake news detection purposed by our team GDUFS_DM in IberLEF 2021 shared task. Our purpose is to use BERT and Sample Memory with an attention mechanism to detect Spanish fake news. To capture richer semantic information in long news texts, we used BERT to encode the news headline and the news beginning and end part to keep more information instead of using a simple truncation strategy. In addition, we also use a matrix parameter initialized by sample representation (we call it Sample Memory), combine with the attention mechanism, our model can capture the relationship information between samples which strengthens the model's robustness in the inference stage. Our submission result achieved the first place on the leaderboard, which fully reflects the advantages of our model.

Keywords: Fake News Detection, Spanish, BERT, Sample Memory.

1 Introduction

Fake news refers to the news articles that are intentionally and verifiably false [1]. The rapid development of online news media platforms not only provided convenience for readers to obtain news information, but also provided soil for the breeding and dissemination of fake news. The publication of fake news is often intentional, and some individuals or organizations may publish different types of fake news for different purposes. Fake news can not only be used to insult and slander individuals, but it can also disrupt social order, instigate political unrest, or even undermine the peace and stability of the international community. What's worse, researches on the dissemination of fake news shows that fake news is significantly faster, deeper, and wider distributed than true news [2]. Therefore, it’s has important practical significance to know how to use the machine to automatically and accurately identify false news.

IberLEF 2021, September 2021, Málaga, Spain.

The FakeDeS@IberLEF 2021 [3, 4] provided us with a fake news detection corpus in Mexican Spanish [5]. The corpus mainly collects Mexican Spanish fake news from websites and contains quite balanced data of real and fake news on 9 different topics, which is intended to encourage people to actively research the identification of fake news in Mexican Spanish to solve the problem of detecting fake news articles in Mexican Spanish spread through digital media. The distribution of the dataset is shown in Table 1. Our team GDUFS_DM also participated in this evaluation and achieved first place on the leaderboard. In this report, we will review our solution for this task, namely, Mexican Spanish Fake News Detection with BERT and Sample Memory (see Section 3.3 for details).

Table 1. The statistics of the Mexican Spanish fake news corpus

	Training Set		Validation Set	
Topic	True	Fake	True	Fake
Economy	18	12	6	7
Education	6	9	4	3
Entertainment	48	55	22	23
Health	16	16	7	7
Politics	121	105	54	43
Science	32	30	14	13
Security	11	18	6	7
Society	41	52	19	22
Sport	45	41	21	17

2 Related work

Numerous scholars had conducted extensive research in text features and emotional features to improve the effect of fake news detection. Ajao et al. pointed out that there is a relationship between the news veracity and the sentiment of the published text and attached a sentiment feature (ratio of the number of negative and positive words) to help the plain text fake news detectors [6]. Instead of attaching a unique feature, Zhang et al. verified the difference between dual emotions in fake news and real news, and proposed a dual emotion feature to represent dual emotions and the relationship between them for fake news detection [7]. Przybyap concluded that the writing style of fake news has certain characteristics, so they designed two new classifiers: a neural network classifier and a model classifier based on stylometric feature [8]. Wang et al. proposed an enhanced weakly supervised fake news detection framework, WeFEND, which can utilize user reports as weak supervision to expand the amount of training data for fake news detection, given the dynamic nature of news and the reality that labeled samples may become outdated quickly [9]. Yi Xie et al. proposed a fake news detection framework to make full use of characterize users by utilizing social user graphs [10].

However, the majority of studies on automated fake news detection have been limited to English documents, and few have evaluated works in other languages. Moreover, the spread of deceptive news tends to be a worldwide problem, so we need to study fake news not only in English, but also look at the world and detect fake news in other languages. Some scholars had also studied fake news detection of some low-resource languages. Nankai Lin et al. proposed the CharCNN-RoBERTa model to detect fake news in the Urdu Language [11]. Hugo Queiroz Abonizio et al. evaluated textual features not linked to a specific language when describing textual data for detecting news [12]. News corpora written in American English, Brazilian Portuguese and Spanish were explored to investigate complexity, stylometric and psychological textual features. As regards the Mexican Spanish, the MEX-A3T@IberLEF2020 [13] has called methods for aggressiveness and fake news detection in Spanish in Mexico. Samuel Arce-Cardenas et al. evaluated the combination of basic text classification techniques, including six machine learning algorithms, two methods for extracting keywords, and two preprocessing techniques [14]. The best results they ran showed an F1-macro score of 0.815 for fake news. Esaú Villatoro-Tello el at. [15] evaluated Supervised Autoencoder (SAE) learning algorithms in a text classification task. They used three different sets of features as input, namely classical word n-grams, char n-grams and Spanish BERT encodings, and obtained the best performance (F = 85.66%) in the fake news classification task.

3 Method

3.1 Overview

The model we finally proposed in this task is shown in Fig. 1. The beginning and end part of the news text is feed into BERT for obtaining two text embeddings (which were called Head Embedding and Tail Embedding). Then after an element-wise addition was applied in those two embeddings, we calculated the dot-product attention between the result and Sample Memory utils to obtain Memory Embedding. Finally, the Beginning Embedding, End Embedding, and Memory Embedding are stitched together to calculate the output result. The specific components will be explained in detail below. In addition, we also use tricks such as gradient accumulating, early stop, and hierarchical learning rates.

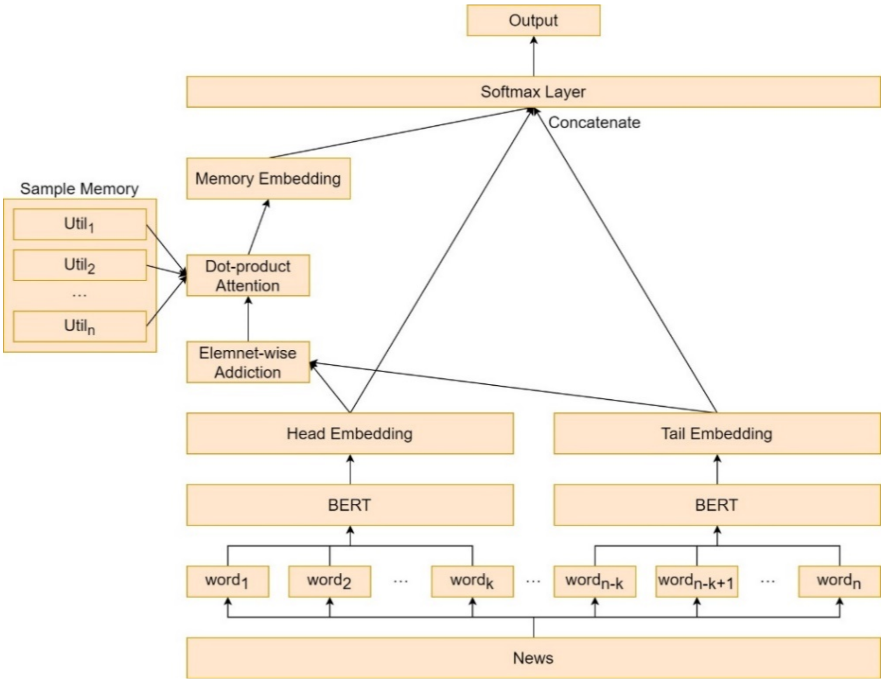

Fig. 1. Fake news detection model we proposed

3.2 BERT

It has been shown that the use of pre-trained language models (PTMs) significantly improves the performance of text classification, and also reduces the amount of labeled sample data required in supervised learning [16]. In this evaluation task, we also used one of the representatives of the pre-trained model, BERT [17] (Bidirectional Encoder Representations from Transformers). In the pre-training section, the model needs to learn the general semantic information of the language from a large-scale unlabeled corpus, according to the pre-training task we set. In fine-tuning section, we can use it as a feature capture model in downstream tasks to obtain the embedding of text or tokens and use different finetune frameworks according to the specific task with labeled data. One disadvantage of the pre-trained model is that pre-training often takes a lot of computing power and time which may be a difficult thing for us. Fortunately, DCC Canete J et al. released the Spanish BERT model on an open-source platform called Transformers [18,19]. This model has two versions of cased and uncased for us to use which were also the main BERT models used in this evaluation.

Most pre-trained models set the maximum sequence length to 512, which is not very friendly to long texts such as news texts. A common solution is truncation, that is, only a sequence of tokens of a limited length is retained. Previous research has found that for text classification, keeping the head and tail tokens at the same time can achieve better results than keeping only the head tokens or only the tail tokens, which means that the head and tail parts of the text may contain more information than the middle part [20]. We concatenated news headlines and news texts in the training set and validation set. After tokenization, we calculated the length distribution of the token sequence (see Fig. 2). The average sequence length reached 546, and nearly 41% of the sequence length is greater than 512. This means that if only a simple truncation method is used, a lot of useful information may be lost. To solve this problem, we adopted a simple method, respectively taking the first 512 tokens and the last 512 tokens for encoding, and concatenating them to get the embedding of the news text to retain as much information as possible.

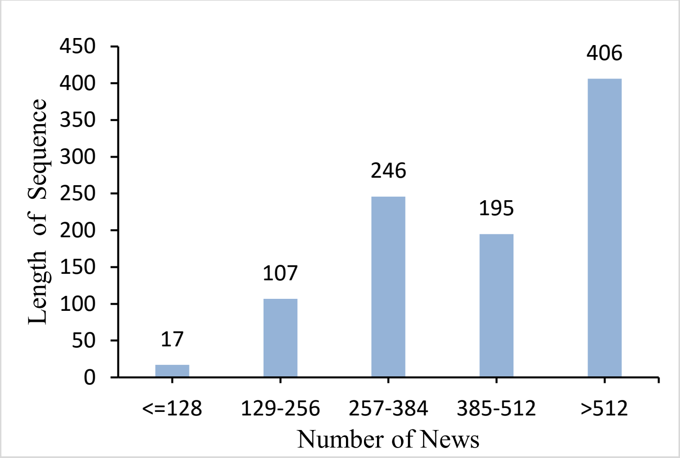

Fig. 2. The length distribution of token sequence

3.3 Sample Memory

This evaluation task also brings two challenges, thematic and language variation. It means that the news in the test corpus may contain topics which are not part of the training corpus, and some test data may be inconsistent with the training corpus in terms of language style. In order to improve the robustness and generalization ability of the model, we also designed a network structure which called "Sample Memory". Sample Memory contains m vectors with the same dimensions as Head Embedding and Tail Embedding (we call it Memory Utils). Before network training, the Head Embedding and Tail Embedding of m samples need to be used for the initialization of Sample Memory. The model accepts a news and applies Dot-product Attention between sample embedding and Sample Memory utils to obtain Memory Embedding. The calculation formula of Memory Embedding is shown in formula 1.

$$E_{memory} = softmax\left(E_{sample} * [U_1{}^T, U_2{}^T, \dots, U_m{}^T]\right) * [U_1, U_2, \dots, U_m]^T \tag{1}$$

Where E_{memory} is Memory Embedding, E_{sample} refers to the element-wise addition result of Head Embedding and Tail Embedding, and $U_1, U_2, \dots, U_m$ respectively refers to m Utils in Sample Memory. Therefore, even if the news text received by the model is very different from the training set during inferencing, Memory Embedding at this time is still not going to change much, which means that the experience learned by the model in the past is still going to be useful. In addition, we concatenate Memory Embedding with Head Embedding and Tail Embedding in the next network layer as usual, the model therefore can better learn the semantic information of the news text itself.

3.4 Tricks

Gradient Accumulation. Limited by the memory size of the training GPU device, the batch size of our model during training can only reach 4. Researchers have found that increasing the batch size appropriately will help the model loss decrease more stably [21]. Therefore, the effect of batch size of 32 is approached by accumulating the gradient of every 8 training steps during model training.

Early stop. Early stop is a Widely used trick in deep learning to avoid overfitting. After each training epoch is finished, we run an evaluation on the validation dataset to obtain validation loss. If the loss value does not continue to decrease within 3 epochs, then the model training will be stopped and the model with the best performance on the validation set will be taken as the model to be submitted.

Differential learning rates. It has been shown that the features captured by different layers in the neural network may be different [22]. Howard and Ruder pointed out that different layers in the neural network should be set to different learning rates according to specific tasks, and the experiment shows that setting different learning rates according to different layers is beneficial to the model to achieve better results [23]. This inspired us to set different learning rates for the parameters of the BERT model, the parameters of the Sample Memory and the parameters of the output layer during the model training process.

4 Experiment

In our experiment, we only used the string concatenated by the news title and body as input. The final result we submitted used the bert-base-spanish-wwm-cased model [15] published by DCC UChile and the spanberta-base model published by Skim AI Technologies [24]. We used 1e-5, 1e-4, and 1e-3 initial learning rates for BERT parameters, Sample Memory parameters, and output layer parameters. Before model training, we selected all training samples, validation samples, and test samples to initialize Sample Memory. In fact, we also tried some machine learning models (using TFIDF for feature extraction) for comparison, such as Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Gradient Boosting Decision Tree (GBDT) and Random Forest (RF). Table 2 shows the results in validation and test set that was reported in accuracy score, precision score for fake (Fake-P), recall score for fake(Fake-R) and f1-score for fake(Fake-F1).

Table 2. Results in validation and test set

	Validation Set				Test Set
Model	Accuracy	Fake-P	Fake-R	Fake-F1	Fake-F1
SVM	74.58	74.45	71.83	73.12	-
NB	54.24	54.02	33.10	41.05	-
LR	73.56	73.53	70.42	71.94	-
DT	65.42	65.38	59.86	62.50	-
GBDT	76.95	78.46	71.83	75.00	-
RF	78.31	74.38	**83.80**	78.81	-
Ours(spanberta-base-cased)	86.10	89.76	80.28	84.76	69.07
Ours(bert-base-spanish-wwm-cased)	**86.44**	**90.48**	80.28	**85.07**	**76.66**
Second best system (in leaderboard)	-	-	-	-	75.48

In order to explore why Sample Memory works in our model, we also designed the following experiment on the validation set: initialize the Sample Memory randomly instead of using news corpus for initialization, and simply remove Sample Memory (see Table 3). It can be seen that even randomly initialized Sample Memory can also improve the performance while using real news samples to initialize Sample Memory can achieve better results. We also found something interesting that the classification precision of fake news has improved after removing Sample Memory, although the recall score has decreased significantly. This may indicate that Sample Memory can guide the model to seek "reference objects" from past samples to guide decision-making. Although errors may sometimes occur, fake news can be more easily detected thanks to Sample Memory.

Table 3. Experiment results for Sample Memory

Model	Fake-P	Fake-R	Fake-F1
Ours(bert-base-spanish-wwm-cased)	90.48	80.28	85.07
Randomly initialized sample memory	89.52(-0.96)	78.17(-2.11)	83.46(-1.61)
Without sample memory	94.23(+3.75)	69.01(-11.27)	79.67(-5.4)

5 Conclusion

In this report, we presented the solution of team GDUFS_DM in the IberLEF 2021 shared task to detect fake news in Spanish. We proposed to use BERT to encode the beginning and end of the news text, and then applied the Sample Memory module let the model to learn the relationship between samples. We also applied some training tricks gradient accumulation, early stop and differential learning rates. The results show that our model got the first place in the ranking. In our future work, we will consider using information such as URLs and topics added to the news to enhance the performance of fake news detection.

6 Acknowledgements

This work was supported by the Key Field Project for Universities of Guangdong Province (No. 2019KZDZX1016), the National Natural Science Foundation of China (No. 61572145) and the National Social Science Foundation of China (No. 17CTQ045). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Shu K, Sliva A, Wang S, et al.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter 19(1), 22-36 (2017).
2. Vosoughi S, Roy D, Aral S.: The spread of true and false news online. Science 359(6380), 1146-1151 (2018).
3. Gómez-Adorno H, Posadas-Durán J P, Bel-Enguix G, and Clau-dia P.: Overview of fakedes task at iberlef 2021: Fake news detection in spanish. Procesamiento del Lenguaje Natural, 67(0), (2021).
4. Manuel Montes, Paolo Rosso, Julio Gonzalo, Ezra Aragón, Rodrigo Agerri, Miguel Ángel Álvarez-Carmona, Elena Álvarez Mellado, Jorge Carrillo-de-Albornoz, Luis Chiruzzo, Larissa Freitas, Helena Gómez Adorno, Yoan Gutiérrez, Salud María Jiménez Zafra, Salvador Lima, Flor Miriam Plaza-de-Arco and Mariona Taulé (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, (2021).
5. Posadas-Durán J P, Gomez-Adorno H, Sidorov G, et al.: Detection of fake news in a new corpus for the Spanish language. Journal of Intelligent & Fuzzy Systems, 2019, 36(5): 4869-4876.
6. Ajao O, Bhowmik D, Zargari S.: Sentiment aware fake news detection on online social networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2507-2511 (2019).
7. Zhang X, Cao J, Li X, et al.: Mining Dual Emotion for Fake News Detection. arXiv e-prints arXiv: 1903.01728 (2019).
8. Przybyla P.: Capturing the Style of Fake News. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34(01), pp. 490-497 (2020).
9. Wang Y, Yang W, Ma F, et al.: Weak supervision for fake news detection via reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34(01), pp. 516-523 (2020).
10. Xie Y, Huang X, Xie X, et al.: A Fake News Detection Framework Using Social User Graph. In: Proceedings of the 2020 2nd International Conference on Big Data Engineering. pp. 55-61 (2020).
11. Lina N, Fua S, Jianga S.: Fake News Detection in the Urdu Language using CharCNN-RoBERTa. In: The 12th edition of the Forum for Information Retrieval Evaluation (FIRE 2020). pp. 447-451. (2020).
12. Abonizio H Q, de Morais J I, Tavares G M, et al.: Language-independent fake news detection: English, Portuguese, and Spanish mutual features. Future Internet 12(5), 87 (2020).
13. Aragón M E, Jarquín H, Gómez M M, et al.: Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican Spanish. In: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain. (2020).

14. Arce-Cardenasa S, Fajardo-Delgadoa D, Álvarez-Carmonab M Á.: TecNM at MEX-A3T 2020: Fake News and Aggressiveness Analysis in Mexican Spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020). pp. 265-272. CEUR Workshop Proceedings (2020).
15. Ramírez-de-la-Rosa G, Parida S, Kumar S, et al.: Idiap and UAM Participation at MEX-A3T Evaluation Campaign. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020). pp. 252-257. CEUR Workshop Proceedings (2020).
16. Peters M E, Neumann M, Iyyer M, et al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018).
17. Devlin J, Chang M W, Lee K, et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
18. Canete J, Chaperon G, Fuentes R, et al.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR (2020).
19. BETO: Spanish BERT, https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased, last accessed 2021/06/01
20. Sun C, Qiu X, Xu Y, et al.: How to Fine-Tune BERT for Text Classification?. arXiv e-prints arXiv: 1905.05583 (2019).
21. You Y, Gitman I, Ginsburg B.: Scaling sgd batch size to 32k for imagenet training. arXiv preprint arXiv:1708.03888 (2017).
22. Yosinski J, Clune J, Bengio Y, et al.: How transferable are features in deep neural networks?. arXiv preprint arXiv:1411.1792 (2014).
23. Howard J, Ruder S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018).
24. Spanish Bert pretrained model released by Skim AI Technologies, https://huggingface.co/skimai/spanberta-base-cased, last accessed 2021/06/01