Automatic Detection of Fake News Spreaders Using BERT

Notebook for PAN at CLEF 2020

Arup Baruah[1], Kaushik Amar Das[1], Ferdous Ahmed Barbhuiya[1], and Kuntal Dey[2]*

[1]IIIT Guwahati, India
[2]Accenture Technology Labs, Bangalore
arup.baruah@gmail.com, kaushikamardas@gmail.com,
ferdous@iiitg.ac.in, kuntal.dey@accenture.com

Abstract This paper discusses the approach we used to detect fake news spreaders. We used the pre-trained large cased BERT model to perform the classification. We experimented by concatenating all the tweets of an author and then performing classification using the vector obtained by max-pooling the 1024-dimensional vectors of the sub-strings of the concatenated string. We also experimented by processing each of the tweets of an author separately. It was found that concatenating the tweets yields better performance. This model obtained an accuracy of 0.6900 on the test set.

1 Introduction

The shared task on "Profiling Fake News Spreaders" was held as part of PAN at CLEF 2020. This task is basically a binary classification task where it is required to determine if a given author has spread some fake news in the past or not. Detecting fake news spreaders is an important step to prevent fake news from propagating through social media. This task was held for English and Spanish languages. The details of this shared task is available in Rangel et al. [8].

In this paper, we describe the work we performed for this shared task. We participated in this task for the English language only. We used the pre-trained large cased BERT [1] model to classify authors as fake news spreaders or not. The rest of this paper is structured as follows: Section 2 discusses the related work that has been performed for author profiling and fake news detection, Section 3 describes the dataset used for this shared task, Section 4 presents the methodology we used, and Section 5 discusses the results we obtained.

2 Related Work

The task of detecting fake news spreaders falls in the category of author profiling. As opposed to author attribution where it is required to determine the identity of the author

* *This work was done when the author was affiliated with IBM Research India, New Delhi

of a particular piece of text, author profiling is about categorizing authors into different classes such as gender, age, occupation, bots, etc. based on a given text as the evidence. PAN has been conducting shared tasks on author profiling since 2013.

Rangel and Rosso in [4] summarizes the author profiling task of PAN 2019. This subtask required determining if the author of a particular piece of text is a human or a bot. If the author is a human, the task also required determining the gender of the author. The best performing system for detecting bots in the English language obtained an accuracy of 0.96 using a random forest classifier [3]. Features used include tweet length, number of capital and lowercase letters, mentions, retweets, edit distance between consecutive tweets, and tf-idf of unigrams and bigrams. For gender classification, the best performance was obtained by a logistic regression classifier [10]. The features used includes word n-grams (1 to 3) and character n-grams (3 to 5). Instead of removing emoji and special characters, they were converted to text. The system obtained an accuracy of 84% in detecting gender. Polignano et al. [5] used a deep learning approach to detect bots and gender. A combination of convolutional neural networks and dense neural networks were used to perform the classification. GloVe, word2vec, and FastText word embeddings were used in the study. Their system obtained accuracy scores of 0.9182 and 0.7973 in detecting bots and gender respectively.

With regard to fake news detection in social media texts, Shu et al. in [9] discuss knowledge-based, style-based, stance-based, and propagation-based approaches for detecting fake news. They also list the different types of features that can be used for fake news detection which include content-based features (source, headline, lexical features, syntactic features, and visual features), and social context features (user-based, post-based, and network-based).

3 Dataset

The shared task "Profiling Fake News Spreaders on Twitter" was conducted for English and Spanish languages. The training data provided for the English language consisted of tweets for 300 different authors and 100 tweets were provided for each author. The dataset was balanced with 150 positive and 150 negative instances. The number of tokens in each tweet varied from 6 to 30.

4 Methodology

In our work, we used the pre-trained large cased BERT model. This version of BERT has 24 layers and 16 attention heads. It produces 1024-dimensional vectors to represent the words. BERT generates contextualized word embeddings as opposed to static embeddings produced by word2vec or GloVe.

The details of our approach are depicted in Figure 1. As mentioned in Section 3, for each author, a list of 100 tweets is provided in the dataset. The tweets for each author were first concatenated. The concatenated string was then tokenized using BERT's WordPiece tokenizer. The tokens were then split into chunks of length 500 tokens. If the last chunk had less than 500 tokens, it was padded with zeroes to make the length equal to 500 tokens. Each of the token sub-list was then provided as input to the pre-trained

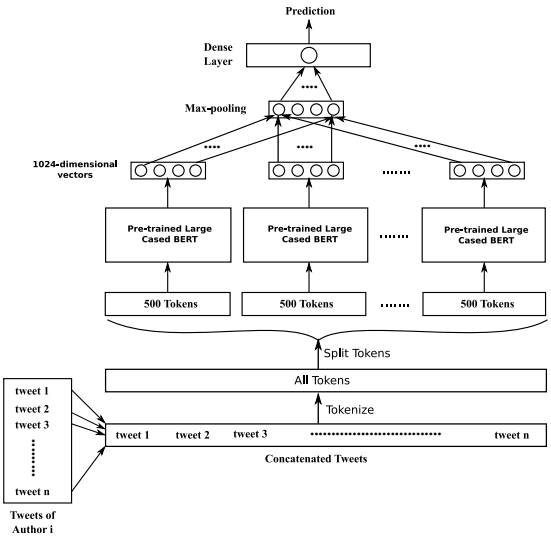

Figure 1. Architecture of our classifier

BERT model. The 1024-dimensional vector from the Extract layer of the BERT model was used as the representation of the sub-string. Max-pooling was then performed on the 1024-dimensional vectors of the token sub-lists. The resultant 1024-dimensional vector was then provided as input to the classification layer. The classification layer consisted of a single Dense layer having a single unit. The sigmoid activation function was used for the layer. The Adam optimizer was used for training and the loss function that was used was is Binary Crossentropy.

We also performed another experiment, whereby, instead of concatenating all the 100 tweets of an author, a 1024-dimensional vector was generated for each tweet using the pre-trained BERT model. Max pooling was performed on the 100 vectors that thus obtained. The classification was performed using the resultant vector. Max sequence length of 40 was used for the experiment.

5 Results

In this section, we discuss the results obtained on the development and the test set. The development set was created from the dataset by doing a stratified split. 80% of the dataset was used for training and 20% of the dataset was used as the development set. Table 1 shows the results obtained on the development set. As mentioned in section 4, experiment 1 in the table refers to the experiment where the tweets of each author were concatenated and then split in sub-lists of 500 tokens each (after tokenization). Experiment 2 in the table refers to the experiment where each tweet was processed separately. As can be seen from the table, concatenating the tweets resulted in better performance than processing each tweet separately. Based on this observation, the model from experiment 1 was used to make the submission for the shared task.

Table 2 shows the confusion matrices for Experiment 1 and Experiment 2 on the development set. As can be seen, Experiment 1 performed better than Experiment 2 in detecting the fake news spreaders category.

Experiment	Precision	Recall	F1	Accuracy
Experiment 1	**0.7229**	**0.7167**	**0.7147**	**0.7167**
Experiment 2	0.6435	0.6333	0.6267	0.6333

Table 1. Dev Set Results

	Experiment 1		Experiment 2	
	Pred NOT	**Pred FAKE**	**Pred NOT**	**Pred FAKE**
NOT	24	6	23	7
FAKE	11	19	15	15

Table 2. Confusion Matrix for Dev Set

System	Method	Accuracy
Our System	BERT	0.6900
Best System	-	0.7500
Baseline 1 [7]	Low Dimensionality Representation	0.7450
Baseline 2	NN + word n-grams	0.6900
Baseline 3	SVM + char n-grams	0.6800
Baseline 4 [2]	LSTM + Emotional features	0.6400
Baseline 5	LSTM	0.5600
Baseline 6	Random	0.5100

Table 3. Test Set Results (English language)

Our model obtained an accuracy score of 0.6900 on the test set for English language. The evaluation on the test set was performed on the TIRA platform [6]. Table 3 shows the performance of our system in comparison to the best performing system of the shared task and other baseline systems. As can be seen, our system performed better than the random, LSTM, emotionally infused LSTM [2], and character n-gram based SVM baseline systems. Our system had the same accuracy as the word n-gram based NN baseline and performed worse than the baseline system that used the low dimensionality representation technique [7]. The final rank in the shared task was determined by averaging the accuracy scores obtained for both English and Spanish languages. As we did not make any submission for Spanish language, our system obtained a rank of 58 out of 66 participants. However, when considering the scores for only English language, we obtained a rank of 29 out of 66 participants.

6 Conclusion

Detecting fake news spreaders in an important step to control the spread of fake news through social media. In our work, we used a classifier based on the pre-trained large cased BERT model to detect fake news spreaders. It was found that concatenating all the tweets of an author yielded a better performance than processing each tweet separately. Our model obtained accuracy score of 0.6900 in the test data. It performed better than the character n-gram based SVM, LSTM, emotionally infused LSTM and the random baseline systems.

References

1. 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. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
2. Ghanem, B., Rosso, P., Rangel, F.: An Emotional Analysis of False Information in Social Media and News Articles. ACM Transactions on Internet Technology (TOIT) **20**(2), 1–18 (2020)
3. Johansson, F.: Supervised classification of twitter accounts based on textual content of tweets. 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_154.pdf
4. Pardo, F.M.R., Rosso, P.: Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling 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_263.pdf
5. Polignano, M., de Pinto, M.G., Lops, P., Semeraro, G.: Identification of bot accounts in twitter using 2d cnns on user-generated contents. 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_95.pdf
6. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. Springer (Sep 2019)
7. Rangel, F., Franco-Salvador, M., Rosso, P.: A Low Dimensionality Representation for Language Variety Identification. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 156–169. Springer (2016)
8. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
9. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. SIGKDD Explorations **19**(1), 22–36 (2017), https://doi.org/10.1145/3137597.3137600

10. Valencia-Valencia, A.I., Gómez-Adorno, H., Rhodes, C.S., Pineda, G.F.: Bots and gender identification based on stylometry of tweet minimal structure and n-grams model. 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_216.pdf