Ensemble of ELECTRA for Profiling Fake News Spreaders

Notebook for PAN at CLEF 2020

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

[1] Indian Institute of Information Technology, Guwahati
{kaushikamardas,arup.baruah}@gmail.com
ferdous@iiitg.ac.in
[2] Accenture Tech Labs, Bangalore, India
kuntal.dey@accenture.com

Abstract This paper presents an ensemble classifier that uses ELECTRA models for the task of identifying possible Fake News Spreaders on Twitter in PAN at CLEF 2020 lab. Our ensemble is created using 15 models which have been fine-tuned on the task dataset. Our approach scored an accuracy of 0.70 and 0.69 on the English and Spanish test sets respectively.

1 Introduction

Fake news is a form of news that is circulated with the aim of deceiving users and manipulating them into formulating specific opinions. With the growth of social media platforms such as Facebook and Twitter, it is now easier than ever to spread fake news. This problem is aggravated further when users knowingly or unknowingly share articles that contain false or misleading information.

There exist numerous sites that use expert analysis to fact check and debunk fake articles, such as snopes.com, politifact.com etc. The problem of fake news has also been actively tackled by the research community. To list a few, the works in [5,6] studied the incorporation of emotional features into Long Short Term Memory (LSTM) network for detecting fake news. The authors in [10] introduced a system called DeClarE which combines evidence collected from the web, language style and trustworthiness of the sources for analysing the credibility of claims in textual form. The work in [17] investigated the use of user profiles as potential features for improving fake news detection systems.

With an aim to further investigate this problem, PAN at CLEF'20 introduced the task of Profiling Fake News Spreaders on Twitter [13]. The objective of this task is to identify whether a Twitter user is a possible fake news spreader given a collection of his tweets. This task is available in English and Spanish. We participated in this task in both languages.

The rest of this paper is organised as follows. First, the dataset for the task is discussed in §2. Our approach is described in §3. The performance of our classifier is analysed in §4 before concluding in §5.

2 Dataset

The dataset [15] for the task of Profiling Fake News Spreaders is featured in two languages: English and Spanish. The number of data samples in each of these two datasets are the same. Each contains 300 authors out of which 150 are labelled as *possible fake news spreaders* while the rest are labelled as *not spreaders*. A collection of 100 tweets is given for each of these authors in which we trained our classification systems. The dataset is perfectly balanced as illustrated in Figure 1. Additionally, the dataset is anonymized [14] to protect the tweet author's privacy. As such, identifiers like user handles and URLs have been replaced with'#USER#' and '#URL#' tokens respectively. Some examples of the data are given in Figure 2. Some other noteworthy features of the dataset are listed below.

- By counting the number of unique tweets within the collection of 100 tweets given for each author, we found that overall, only 343 authors have all unique tweets. For each language, about half of the authors had some duplicates.
- In the entire dataset, the shortest tweet has 1 word while the longest tweet has 86 words. The tweets contained an average of around 15 words.
- While many authors did not use any emojis, 284 to be exact, the rest used emojis in at least one of their tweets.

We do not know anything about the test set since our models were evaluated using the TIRA system [11]. TIRA uses *blind evaluation*, a paradigm in which it runs our models on a hidden test set without exposing any information about it to the participants. We submitted our models adhering to the guidelines given by the organizers.

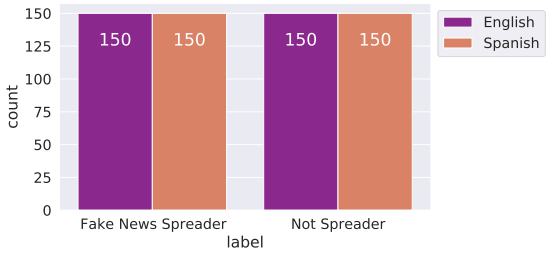

Figure 1. Data distribution

English: 'Journaling Benefit: How Journaling Can Help Create Mental Calmness and Clarity #URL# #HASHTAG#... #URL#'
Spanish: 'Abuelito pagará 2 mil pesos por daños al vehículo que lo atropelló #URL#'

Figure 2. Example tweets from the dataset

3 Methodology

Our approach involves an ensemble classifier built using the ELECTRA model [3]. We chose this model because of it's small size, fast training speed and promising benchmark scores. This made it possible for us to experiment with ensembles using modest computing resources. In this section, we briefly describe the ELECTRA model before moving on to the details about the classifier. In the rest of this section, *author* and *data sample* are used interchangeably, since each data sample in the dataset is an author. The code used in this work is available in GitHub[3].

3.1 ELECTRA

At present, the current state-of-the-art in natural language processing is held by large Transformer-based [18] models which have been trained using the BERT technique [4], for example, RoBERTa [8], T5 [12], etc. These models are trained in an unsupervised manner on a vast amount of text data and can be fine-tuned for other downstream tasks such as text classification, question answering, etc. The unsupervised task often used for training such models is the prediction of masked tokens. In this task, a small percentage, typically 15%, of tokens in the input data is corrupted with a [MASK] token. The model is trained to correctly predict these masked tokens. This task (also called as a *pre-training task*) has the disadvantage of only learning from a small portion of the text sequence which is largely computationally inefficient.

The ELECTRA training method [3], which stands for *Efficiently Learning an Encoder that Classifies Token Replacements Accurately*, aims to address this inefficiency while retaining all the same capabilities of BERT. This is done by using a novel pre-training task, called as *replaced token detection*, in which a model is trained to distinguish between real input tokens from synthetic but plausible replacements. The model is required to predict over each of the input tokens whether it is the real input token or a replaced one thereby learning from the entire sequence instead just a small percentage of it. This results in ELECTRA performing competitively with other state-of-the-art-models while using only about 25% of their computing requirements.

3.2 Token limit of ELECTRA

Most BERT-style transformers have a token limit of 512. Models trained using ELECTRA have the same limitations. This makes it difficult to directly feed the given data samples into our ELECTRA based classification system. It is because for each author,

[3] https://github.com/cozek/profiling-fake-news-spreaders

i.e for each data sample, a collection of 100 tweets is given, whose tokens altogether cross the token limit.

An obvious method would be to truncate and reduce the number of tokens. But we avoid doing so due to two reasons. Firstly, in the entire set of an author's tweets, not all of them might be fake and vice versa. Secondly, doing so will result in the loss of a lot of information. Hence, to address these issues, in this work, we randomly sample n tweets from an author's set of tweets. The exact implementation details and intuition is explained in §3.3.

3.3 Random Sampling of an Author's Tweets

Intuition The intuition behind random sampling is that we do not know which of the tweets from an author's set of tweets are relevant for the classification task. So, at each training epoch, if we randomly sample an author's tweets to feed into the model, the model will have the chance to look at enough of an author's tweets to learn if the author is a fake news spreader or not.

Implementation While constructing a batch of samples to feed into the model, from each author, randomly n tweets from the collection of 100 are selected. These are then concatenated with special classification tokens as given in Figure 3. This chosen collection of random tweets for each author is not fixed and is randomly chosen again at every epoch. Therefore, tweets chosen in a previous epoch might get chosen again. We use $n = 14$ so that the token limit is never exceeded even in edge cases where the tweets might be longer.

If T^a is the set of tweets of an author a, t is a subset of randomly selected n tweets from T^a at the i^{th} training epoch such that $n \leq |T^a|$, then C_i^a is the concatenation of the tweets in t, where C_i^a is defined as

$$C_i^a = \text{<s>}\ t_1\ \text{<\s>}\ t_2 ...\ \text{<\s>}\ t_{n-1}\ \text{<\s>}\ t_n \quad (1)$$

Here is <s> and <\s> are special tokens defined in ELECTRA's vocabulary as CLS_TOKEN and SEP_TOKEN respectively. CLS_TOKEN marks sentences for classification. SEP_TOKEN separates each tweet.

Figure 3. Concatenation Method

3.4 Ensemble Classifier

One obvious drawback of random sampling described in §3.2 is that looking at only a small random portion of an author's tweets may not enough to make a correct decision. To mitigate this problem, we use an ensemble. Our proposed ensemble is built using 15 fine-tuned models each of which is built on top of a pre-trained ELECTRA model.

Each model of the ensemble looks at a *different* random sample of an author's tweets and makes a prediction. The final prediction is determined by *majority voting* where the label with the highest frequency is chosen as the final label for the task. This ensures that a wide range of an author's tweets is looked at before coming to a decision. Ensembling also has the effect of lowering the variance of the model [16]. The architecture of the models in the ensemble and the training routine is described below.

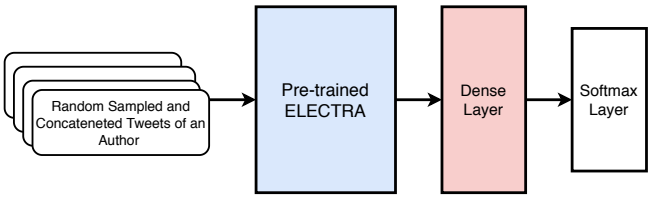

Figure 4. A single classification model of the ensemble. The dense layer is $\tanh$ activated and has a dropout of 0.1. The $softmax$ layer makes the prediction.

Model Architecture The ensemble is made of 15 fine-tuned models each of which is of the architecture given in Figure 4. We use only 15 models because adding more does not improve the ensemble [16]. In each model, 256-dimensional embeddings produced by pre-trained ELECTRA are fed into a $\tanh$-activated dense layer having 256 in-features and 256 out-features. After applying a dropout of 0.1 on the output of the dense layer, the output representation is fed into a $softmax$ layer which makes the prediction. The weights of the dense layer and $softmax$ layer are *randomly initialized* in each of the models of the ensemble.

Separate ensembles were built for the two languages in the dataset, one ensemble for English and one ensemble for Spanish. For English, we used the pre-trained model called *google/electra-small-discriminator* from the HuggingFace Transformers Library [4] [19]. Since no official pre-trained model was available for Spanish, we used a pre-trained model called *skimai/electra-small-spanish* from HuggingFace community models hub.

Model Training and Inference The same training routine is applied to each of the models in the ensemble. Each model is fine-tuned with a small learning rate of $\approx 1e-3$ using a cross-entropy loss function for 20 epochs. 90% of the data is used as the train set and the remaining 10% is used as the validation set. The percentage of each class is preserved in both of these. Early stopping was used to stop training if validation accuracy did not improve for 4 consecutive epochs. The model is optimized using Ranger optimizer, which is a combination of LookAhead [20] and RAdam [7]. The (α, k) parameters of the optimizer are set to $(0.5, 5)$. During both training and inference, the random sampling approach described in §3.3 is used to feed data into the model. Data

[4] https://huggingface.co

is fed into the models in batches of 50. As mentioned in §3.4, the final label is determined by majority voting.

4 Results

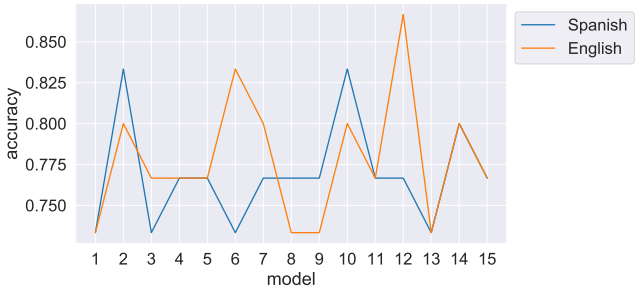

Figure 5. Validation accuracy of each model in ensemble

The accuracy on the validation set of each model in the ensemble is plotted in Figure 5. As apparent from the plot, the accuracy of each model in the ensemble varies widely. This is expected since each model started off with different randomly initialized weights and trained on randomly sampled data. The results obtained by running the ELECTRA ensembles on the validation set and test set is given in Table 4. The ensembles scored an accuracy of 0.70 and 0.69 on the English and Spanish test sets respectively as given in Table 4. Both of the ensembles had almost the same accuracy on the test set. The ensemble for English had a validation set accuracy of 0.87 while the ensemble for Spanish had a validation set accuracy of 0.77. This suggests that the ensembles suffered from over-fitting since there is a notable difference between the validation and test set accuracy. Another explanation could be that random sampling did not sample the relevant tweets for the classifier to be able to differentiate correctly.

	Accuracy	
Language	*Validation Set*	*Test Set*
English	0.87	0.70
Spanish	0.77	0.69

Table 1. Results

5 Conclusion

This paper explored the application of ensembled ELECTRA models for the task of Profiling Fake News Spreaders in Twitter. Random sampling was used in an attempt to

overcome the limitation of the max number of tokens supported by transformer models. In future work, it would be interesting to explore transformer models that do no have such limitations, for example the Longformer [2]. Also, the study did not make use of the many features of the data. We found that the data had duplicates (see §2) that could have been removed during preprocessing. This perhaps might have improved the classifier's performance by preventing duplicates from being sampled. Another promising avenue for future work would to enhance the proposed classifier with emotional signals from the text using lexicons such as EmoLex [9] and SentiSense [1].

References

1. de Albornoz, J.C., Plaza, L., Gervás, P.: Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In: LREC (2012)
2. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv:2004.05150 (2020)
3. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
5. Ghanem, B., Rosso, P., Rangel, F.: An emotional analysis of false information in social media and news articles. ACM Trans. Internet Technol. 20(2) (Apr 2020), https://doi.org/10.1145/3381750
6. Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 877–880. SIGIR'19, Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3331184.3331285
7. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265 (2019)
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
9. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. ArXiv abs/1308.6297 (2013)
10. Popat, K., Mukherjee, S., Yates, A., Weikum, G.: DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 22–32. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://www.aclweb.org/anthology/D18-1003
11. 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)
12. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2019)
13. 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)

14. Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Language and Law= Linguagem e Direito 5(2), 95–117 (2019)
15. Rangel, F., Rosso, P., Ghanem, B., Giachanou, A.: Profiling fake news spreaders on twitter (Feb 2020), https://doi.org/10.5281/zenodo.3692319
16. Risch, J., Krestel, R.: Bagging BERT models for robust aggression identification. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. pp. 55–61. European Language Resources Association (ELRA), Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.trac-1.9
17. Shu, K., Wang, S., Liu, H.: Understanding user profiles on social media for fake news detection. 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) pp. 430–435 (2018)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv abs/1910.03771 (2019)
20. Zhang, M.R., Lucas, J., Hinton, G., Ba, J.: Lookahead optimizer: k steps forward, 1 step back (2019)