TIB's Visual Analytics Group at MediaEval '20: Detecting Fake News on Corona Virus and 5G Conspiracy

Gullal S. Cheema[1], Sherzod Hakimov[1], Ralph Ewerth[1,2]
[1]TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]L3S Research Center, Leibniz University Hannover, Germany
{gullal.cheema,sherzod.hakimov,ralph.ewerth}@tib.eu

ABSTRACT

Fake news on social media has become a hot topic of research as it negatively impacts the discourse of real news in the public. Specifically, the ongoing *COVID-19* pandemic has seen a rise of inaccurate and misleading information due to the surrounding controversies and unknown details at the beginning of the pandemic. The *FakeNews* task at MediaEval 2020 tackles this problem by creating a challenge to automatically detect tweets containing misinformation based on text and structure from Twitter follower network. In this paper, we present a simple approach that uses BERT embeddings and a shallow neural network for classifying tweets using only text, and discuss our findings and limitations of the approach in text-based misinformation detection.

1 INTRODUCTION AND RELATED WORK

The *FakeNews* task [16][1] focuses on automatically predicting whether a tweet consists of misinformation (conspiracy) over the use of two concepts *COVID-19* and *5G* network. The dataset also consists of other conspiracy tweets that are either over some other concepts or accidentally contain the two buzzwords. The challenge requires the participants to mainly develop text or structure based detection models to automatically detect conspiracy tweets.

In the last five years, social media fake news detection has attracted a lot of research interest in academia and industry. Consequently, the problem has been approached from different perspectives including stance detection [10, 18], claim detection and verification [3, 8], sentiment analysis [2, 6], etc. To learn a model from text, recently different variants of neural networks have been used for fake news detection. Convolutional Neural Networks (CNN) in general have been extensively used with word embeddings in several works [9, 13, 20] for social media fake news detection. Recently, Ajao *et. al.* [1] proposed a hybrid model with a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) to identify fake news on Twitter. In similar CLEF challenges [3, 8] over the years, the problem of claim detection in tweets has been tackled with a combination of rich set of features and different kinds of classifiers like SVM [22], gradient boosting [21] and sequential neural networks [11]. However, several works [5, 19] recently have moved from using word2vec [14] type embeddings to rich contextual deep transformer *BERT* (Bidirectional Encoder Representations from Transformers) [7] like embeddings.

[1]https://multimediaeval.github.io/editions/2020/tasks/fakenews/

MediaEval'20, December 14-15 2020, Online

2 APPROACH

We approach the problem from only the textual perspective and rely on training a shallow neural network over contextual word embeddings. Our submitted models use the recently proposed *BERT-large* based model pre-trained [15] on a large corpus of *COVID* Twitter data. This essentially improves the performance by 3-4% in comparison to vanilla *BERT* (V-BERT) since the embeddings are better aligned (with regard to *COVID*) for the task at hand. We also experiment with additional features like sentiment, subjectivity and lexical features that have been shown to improve performance in similar tasks [3, 8]. We observed no improvements using combination of these features and excluded them in this paper.[2]

Text Pre-processing For vanilla *BERT-large*, we use Baziotis *et. al.*'s [4] tool to apply the following normalization steps: tokenization, lower-casing, removal of punctuation, spell correction, normalize *hashtags, all-caps, censored, elongated and repeated* words, and remove terms like *URL, email, phone, user mentions*. For *COVID* Twitter BERT [15], we follow their pre-processing which normalizes text, and additionally replaces *user mentions, emails, URLs* with special keywords.

Contextual Feature Extraction To get one embedding per tweet, we follow the observations made by Devlin et al. [7] that different layers of *BERT* capture different kinds of information, so an appropriate pooling strategy should be applied depending on the task. The paper also suggests that the last four hidden layers of the network are good for transfer learning tasks and thus we experiment with 4 different combinations, i.e., concatenate last 4 hidden layers (4-CAT), the average of last 4 hidden layers (4-SUM), last hidden layer (LAST), and 2^{nd} last hidden layer (2-LAST). We normalize the final embedding so that $l2$ norm of the vector is 1.

Shallow Neural Network We use the extracted and pooled *BERT* embeddings and train a two-layer neural network. Before passing the features through the first layer, we apply a squeeze and excitation (SE) operation [12] that enhances the representation to learn a better model. The SE operation has been shown to improve feature representations and performance in CNNs. Then, the embedding is projected down to 128 dimensions, which is followed by batch normalization and ReLU operation to introduce non-linearity. The 2^{nd} layer is a linear classification layer that produces a softmax probability for each class. Dropout with rate of 0.2 and 0.5 is applied after SE operation and first layer to avoid over-fitting.

Final Prediction on Test Set We train five models on 5-fold splits and take the majority label as the predicted label for the tweet. As described in the challenge, the 3-class submissions can have an additional *cannot-determine* class. We assign a tweet with this label

[2]Source code: https://github.com/cleopatra-itn/TIB_VA_MediaEval_FakeNews

if the softmax probability is less than 0.4, which signifies that the model is not confident enough.

3 EXPERIMENTAL RESULTS

The *FakeNews* development set consists of 5,999 extracted tweets over three classes: *5G* and *COVID-19* conspiracy (1,128 samples), Non-conspiracy (4,173 samples) and other conspiracy (698 samples). We generate five-fold stratified training and validation splits in the ratio of 80:20, so that the distribution of classes remains the same in the training and validation sets. The official test data originally consisted of 3,230 tweets, out of which 308 are not valid or non-existent at the time of evaluation since participants were required to crawl tweets on their own. Table 1 shows the performance of different models on validation sets, while Table 2 shows the evaluation of our submitted models on the official test data. All the runs for the test data use *COVID* Twitter BERT (C-BERT) extracted features. Official metric is Matthews correlation coefficient (MCC). For validation sets, we provide both average accuracy (ACC) and MCC scores. In Table 1, we also show the result of fine-tuning (FT) the last two and four layers of 2 BERT variants with a linear classification layer on top of BERT's *CLS* embedding. Although the average performance of finetuning 4 layers is marginally better than the fixed average word embeddings, the highest in two of the splits is better in the fixed embedding plus neural network.

4 DISCUSSION

Our findings and observations from the FakeNews task can be summarized as follows:

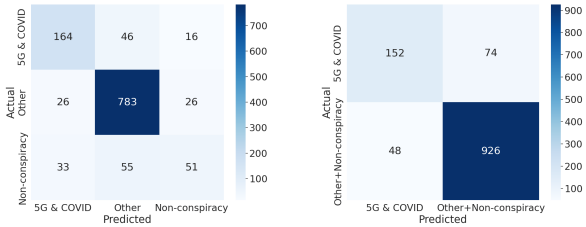
- Vanilla *BERT* clearly has a wider domain gap to perform well on this task, as the concepts and keywords related to *COVID* are fairly recent. The *COVID* Twitter *BERT* outperforms in our finetuning experiment as well as with a shallow neural network on the extracted embeddings.
- The pooling operation and the number of last layers to obtain a sentence embedding does make a difference, as only using the last layer (or 2^{nd} last) performs marginally lower across the metrics. An even better embedding could be a sentence embedding extracted from a sentence-transformer [17], but only if it is pretrained on a *COVID* Twitter corpus to narrow down the domain and knowledge gap.
- Although two-class prediction performance has higher metric scores, merging the other-conspiracy and non-conspiracy tweets decreases the true positives (see Figure 1) for the conspiracy class. This could be because the model is able to learn and detect the conspiracy aspect in tweets, and merging the other two categories negatively impacts the learning.
- In similar social media challenges, pre-processing text also plays a significant role. Therefore, we experimented with replacing different keywords like *corona, sars cov2, wuhan virus, ncov, korona, koronavirus* with *coronavirus* or *covid*, and similarly *five g, fiveg, 5 g* with *5g*. Unfortunately, doing so degraded the performance in some splits and was not a part of our submission model.

Table 1: Model Evaluation of Validation Set.

Model	Pooling	Classes	ACC	MCC
V-BERT (FT)^2	*CLS*	3	0.7616 ± 0.0125	0.4521 ± 0.0347
V-BERT (FT)^4	*CLS*	3	0.7656 ± 0.0094	0.4611 ± 0.0266
C-BERT (FT)^2	*CLS*	3	0.8131 ± 0.0063	0.5837 ± 0.0107
C-BERT (FT)^4	*CLS*	3	**0.8171 ± 0.0091**	**0.5952 ± 0.0192**
C-BERT + NN	4-SUM	3	0.8138 ± 0.0112	0.5793 ± 0.0272
	4-CAT	3	**0.8163 ± 0.0112**	**0.5841 ± 0.0264**
	2-LAST	3	0.813 ± 0.0073	0.5761 ± 0.0181
	LAST	3	0.8083 ± 0.0085	0.5662 ± 0.0211
C-BERT + NN	4-SUM	2	0.8881 ± 0.0122	0.6173 ± 0.0401
	4-CAT	2	**0.8901 ± 0.0087**	**0.6236 ± 0.0330**
	2-LAST	2	0.8886 ± 0.0103	0.6184 ± 0.0357
	LAST	2	0.8838 ± 0.0058	0.5976 ± 0.0218

Table 2: Official Test Data Results. 3-class submissions include an additional *cannot-determine* (CD) class when the model confidence is lower than 0.4.

Submission	Classes	MCC
Run 1 (4-SUM)	3+CD	0.5717
Run 2 (4-CAT)	3+CD	0.5773
Run 3 (4-SUM)	2	0.5986
Run 4 (4-CAT)	2	0.6083



(a) 3-class **(b) 2-class**

Figure 1: Confusion matrices for the different settings

5 CONCLUSION

In this paper, we have presented our solution for the FakeNews detection task of MediaEval 2020. The described solution is based on extracting embeddings from transformer models and training shallow neural networks. We compared the two transformer models and observed that *BERT* transformer pre-trained on *COVID* tweets performs better than vanilla version. Pooling operations such as concatenation or averaging of embeddings of the last hidden layers also play an important role as shown by experimental evaluation. In future work, we will focus on the integration of additional contextual information that is presented via external links along with data from other modalities such as images.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no 812997 (CLEOPATRA ITN).

REFERENCES

[1] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2018. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th international conference on social media and society*. 226–230.

[2] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2507–2511.

[3] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, and others. 2020. Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 215–236.

[4] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 747–754.

[5] Gullal S. Cheema, Sherzod Hakimov, and Ralph Ewerth. 2020. Check_square at CheckThat! 2020 Claim Detection in Social Media via Fusion of Transformer and Syntactic Features. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020 (CEUR Workshop Proceedings)*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol (Eds.), Vol. 2696. CEUR-WS.org. http://ceur-ws.org/Vol-2696/paper_216.pdf

[6] Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. SAME: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 41–48.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the CLEF-2019 CheckThat! Lab: automatic identification and verification of claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 301–321.

[9] Yong Fang, Jian Gao, Cheng Huang, Hua Peng, and Runpu Wu. 2019. Self multi-head attention-based convolutional neural networks for fake news detection. *PloS one* 14, 9 (2019), e0222713.

[10] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180* (2018).

[11] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma. 2018. The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 CheckThat! Lab.. In *CLEF (Working Notes)*.

[12] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[13] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. 2020. FNDNet–A deep convolutional neural network for fake news detection. *Cognitive Systems Research* 61 (2020), 32–44.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013), 3111–3119.

[15] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv preprint arXiv:2005.07503* (2020).

[16] Konstantin Pogorelov, Daniel Thilo Schroeder, Luk Burchard, Johannes Moe, Stefan Brenner, Petra Filkukova, and Johannes Langguth. 2020. FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020. In *MediaEval 2020 Workshop*.

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3973–3983.

[18] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv preprint arXiv:1707.03264* (2017).

[19] Evan Williams, Paul Rodrigues, and Valerie Novak. 2020. Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of Claims using Transformer-based Models. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020 (CEUR Workshop Proceedings)*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol (Eds.), Vol. 2696. CEUR-WS.org. http://ceur-ws.org/Vol-2696/paper_226.pdf

[20] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. TI-CNN: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749* (2018).

[21] Khaled Yasser, Mucahid Kutlu, and Tamer Elsayed. 2018. bigIR at CLEF 2018: Detection and Verification of Check-Worthy Political Claims.. In *CLEF (Working Notes)*.

[22] Chaoyuan Zuo, Ayla Karakas, and Ritwik Banerjee. 2018. A Hybrid Recognition System for Check-worthy Claims Using Heuristics and Supervised Learning.. In *CLEF (Working Notes)*.