Classifying COVID-19 Conspiracy Tweets with Word Embedding and BERT

Yuta Yanagi[1], Ryohei Orihara[1], Yasuyuki Tahara[1],
Yuichi Sei[1], Akihiko Ohsuga[1]
[1]The University of Electro-Communications, Japan
yanagi.yuta@ohsuga.lab.uec.ac.jp,orihara@acm.org,tahara@uec.ac.jp,sei@is.uec.ac.jp,ohsuga@uec.ac.jp

ABSTRACT

We, the team OTS-UEC contributed the automatic detection of conspiracy tweets in MediaEval 2021. The dataset has tweets that refer to COVID-19. Part of them argues/discusses the relationship of conspiracies. Following the results of the MediaEval 2020 working notes, we use a BERT-based classifier. We implement three proposed models and compare them in the experiments. In the task of this year, the model also shows better results of classifying than a text embedding-based one. This result suggests that using the pre-trained model is also suitable to classify conspiracy tweets by small preparation processes.

1 INTRODUCTION

FakeNews, one of the MediaEval 2021 tasks focuses on the automatic classifying of tweets by conspiracies [10]. The FakeNews has three classification subtasks. The first (Text-Based Misinformation Detection, MD) is classifying three stances classes. The given three labels are supporting, discussing, and non-conspiracy (not mention conspiracy). The second (Text-Based Conspiracy Theory Recognition, CTR) is nine binary classifications for pre-defined conspiracies if referred to or not. The third (Text-Based Combined Misinformation and Conspiracies Detection, CMCD) requires classifying three stances by the nine conspiracies (3×9 output types).

We compared the effect of using pre-trained language models in every subtasks. In addition, we attempt to compare two language models. One is pre-trained NNLM [2] based, another is pre-trained BERT [4] based. The results show there are solid improvements in using the pre-trained language model. Moreover, using the BERT based model gives the best result in the experiments.

2 RELATED WORK

The epidemic of COVID-19 affects not only in medical area but also social media. Diffusing misinformation (including fake news) reduces the credibility of governments and medical treatments like vaccines [13]. Moreover, part of people argues the relationships between the epidemic and conspiracies by psychological influences [5, 14]. Therefore, the automatic detection of conspiracy tweets is crucial to lighten the burden imposed on medical workers.

The FakeNews task in 2021 extends from the automatic detection of the 5G conspiracy from COVID-19 tweets in MediaEval 2020 [11, 12]. Among its participants, two teams used the BERT model in a single model [8] or an ensemble model [9]. In both cases, using the BERT model improved classification performance for the 5G conspiracy/the other conspiracy/the non-conspiracy.

MediaEval'21, December 13-15 2021, Online

3 APPROACH

In this section, we show how to implement our proposed model in each subtask.

3.1 Preprocessing

The organizer sent us raw tweet texts as a dataset. Therefore, we apply to preprocess following rules.

- Fix contracted forms by a provided tool [15] and manual processes.
- Make all alphabets to lowercase.
- Remove letters except for alphabets, numbers, and whitespaces.
- Replace all numbers to zero (0) except "covid19"
- Eliminate stopwords by a tool from NLTK [3].

The removed letters include emojis. When we improve the performances of the classifications, considering emojis may be able to extract more accurate tweet features.

3.2 Language Models

The FakeNews task requires making two model types. On the one hand, "required run" needs to complete within the dataset. On the other hand, "optional run (s)" allows using data outside the dataset. The outside data includes pre-trained language models.

We compare the effect of pre-trained language models on the difference of results between these two model types. We have done all implementations in Keras [7].

3.2.1 Required run. First of all, we get encoded tweets consisting of integers by `TextVectorization`. Secondly, we obtain word embedding by the `Embedding` layer. We initialize the layer by the uniform distribution. Finally, we obtain a tweet feature by average-pooling of all word embeddings in `GlobalAveragePooling1D`. The dimensionality of output from the pooling is 128. We add a fully connected layer with a 10% dropout layer. The 32-dimensional array is the tweet features in the required run.

3.2.2 Optional runs. In this run, we can use outside the given dataset includes pre-trained models. We use the BERT-based language model from the results of the FakeNews task in MediaEval 2020 [8, 9]. We assign `small_bert` from TensorFlow Hub [6]. We also add the fully connected layer and obtain the tweet features. In the stance classification subtask, we also compared with a NNLM based language model by TensorFlow Hub [2].

Figure 1: Comparison table between the given labels and the new labels in Misinformation Detection.

Table 1: Results of implemented models.

Subtask name	Model	MCC
Misinformation Detection	Word emb.	0.142
	BERT-based	0.413
	NNLM-based	0.388
Conspiracy Theory Recognition	Word emb.	0.133
	BERT-based	0.267
Combined Misinformation & Conspiracies Detection	Word emb.	0.000
	BERT-based	0.000

3.3 Classification Models

We prepare three classification models for each subtask. We input the tweet features for them.

3.3.1 Misinformation Detection. We build two binary classifiers because the ratio of the labels is nearly 2:1:1. Figure 1 shows the correspondences of the given labels and ones in this subtask imposed by us. The first one considers if a tweet refers to any conspiracies. If it does, the second one considers if the tweet supports the conspiracies or not. Therefore, during the training sequence, the non-conspiracy tweets are not used for the second classifier. We think this will help to train without bias from the imbalance of given labels. We compare in experiments the effect of this structure with the model that classified directly for three labels.

3.3.2 Conspiracy Theory Recognition. We build a classifier for nine outputs that parallel pre-defined conspiracies. We use another fully connected layer that outputs nine values.

3.3.3 Combined Misinformation and Conspiracies Detection. We prepare nine three-class classifiers that deduce stances. We do not use two binary classifiers due to the lack of tweets that refer to each conspiracy.

4 RESULTS AND ANALYSIS

4.1 Effect of Language Model

Table 1 shows the returned results of the FakeNews task. All result values are the Matthews correlation coefficient (MCC) [1].

We can confirm that using the language model makes the results improve except for the CMCD subtask. In the CMCD, all output labels are one in those models, which means non-conspiracy. We attribute this to the fact that by separating the classifiers by the pre-defined conspiracies, we increased the ratio of non-conspiracy tweets. According to the task organizer, other participants also send the all-one output.

Table 2: Detail results of MCC in the conspiracy detection. See also overview paper for every abbreviation of pre-defined conspiracies [10].

Model	SC	BMC	A	FV	IP	HRI	PR	NWO	S
Emb.	0.01	0.16	0.30	-0.09	**0.15**	0.1	0.19	0.2	**0.16**
BERT	**0.04**	**0.41**	**0.44**	0.09	0.05	**0.53**	**0.30**	**0.41**	0.13

Table 3: The results of the couple binary classifications and single three-class classification with BERT in the MD.

Model name	MCC
Double binary classifications	**0.413**
Single three-class classification	0.258

Table 2 shows the detailed result in the CTR subtask. The BERT model is better in seven of the nine pre-defined conspiracies. Even in the remaining three cases the differences are tiny.

4.2 Double Binary Classification

Table 3 the results of two classification structures. Both of them use the BERT-based language models. We can confirm that the double binary classification shows a better score than the single three-class one.

5 DISCUSSION AND OUTLOOK

In this paper, we participate the FakeNews task that requires classifying tweets by conspiracies. To realize it, we employ pre-trained language models from other models for the FakeNews task of MediaEval 2020 [11]. We compare them with models that use only word embedding. According to the experimental result, the pre-trained language model help to extract conspiracy information at the stance classification and the conspiracy detection. However, in classification for the CMCD subtask, all output scores are the same label. We guess that the classification models do not work because the tweets mentioning each pre-defined conspiracy are scattered. However, looking at the models of other teams, it is possible that we have designed our models incorrectly for the CMCD subtask. A closer look at the result of CTK shows variation in the effectiveness of the pre-trained language model by the pre-defined conspiracies. This result may come from the characteristics of the trend of tweet content. It can be needed further researching. Moreover, we also compare the two classification structures at the MD subtask. The experiment results show us that the double binary classification is better than the single three-class classification. We expect this reason is nearly 2:1:1 of three classes ratio. If the ratio is different, the trend will not continue.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP18H03229, JP18H03340, 18K19835, JP19H04113, JP19K12107, JP21H03496.

REFERENCES

[1] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: an overview . *Bioinformatics* 16, 5 (05 2000), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The journal of machine learning research* 3 (2003), 1137–1155.

[3] Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Barcelona, Spain, 214–217. https://www.aclweb.org/anthology/P04-3031

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019). arXiv:cs.CL/1810.04805

[5] Karen M. Douglas. 2021. COVID-19 conspiracy theories. *Group Processes & Intergroup Relations* 24, 2 (2021), 270–275. https://doi.org/10.1177/1368430220982068 arXiv:https://doi.org/10.1177/1368430220982068

[6] Tensorflow hub. 2021. small_bert/bert_en_uncased_L-4_H-512_A-8. (2021). https://tfhub.dev/tensorflow/small

[7] Nikhil Ketkar. 2017. Introduction to keras. In *Deep learning with Python*. Springer, 97–111.

[8] Andrey Malakhov, Alessandro Patruno, and Stefano Bocconi. 2020. Fake News Classification with BERT. In *Working Notes Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020 (CEUR Workshop Proceedings)*, Steven Hicks, Debesh Jha, Konstantin Pogorelov, Alba García Seco de Herrera, Dmitry Bogdanov, Pierre-Etienne Martin, Stelios Andreadis, Minh-Son Dao, Zhuoran Liu, José Vargas Quiros, Benjamin Kille, and Martha A. Larson (Eds.), Vol. 2882. CEUR-WS.org. http://ceur-ws.org/Vol-2882/paper38.pdf

[9] Olga Papadopoulou, Giorgos Kordopatis-Zilos, and Symeon Papadopoulos. 2020. MeVer Team Tackling Corona Virus and 5G Conspiracy Using Ensemble Classification Based on BERT. In *Working Notes Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020 (CEUR Workshop Proceedings)*, Steven Hicks, Debesh Jha, Konstantin Pogorelov, Alba García Seco de Herrera, Dmitry Bogdanov, Pierre-Etienne Martin, Stelios Andreadis, Minh-Son Dao, Zhuoran Liu, José Vargas Quiros, Benjamin Kille, and Martha A. Larson (Eds.), Vol. 2882. CEUR-WS.org. http://ceur-ws.org/Vol-2882/paper76.pdf

[10] Konstantin Pogorelov, Daniel Thilo Schroeder, Stefan Brenner, and Johannes Langguth. 2021. FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021. In *the MediaEval 2021 Workshop, Online, 13-15 December 2020*. Online.

[11] 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*. Online.

[12] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *2021 Workshop on Open Challenges in Online Social Networks*. Online, 21–25.

[13] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science* 7, 10 (Oct. 2020), 201199. https://doi.org/10.1098/rsos.201199

[14] Joseph E Uscinski, Adam M Enders, Casey Klofstad, Michelle Seelig, John Funchion, Caleb Everett, Stefan Wuchty, Kamal Premaratne, and Manohar Murthi. 2020. Why do people believe COVID-19 conspiracy theories? *Harvard Kennedy School Misinformation Review* 1, 3 (2020).

[15] Pascal van Kooten. 2020. contractions. https://github.com/kootenpv/contractions. (2020).