Detecting Fake News Conspiracies with Multitask and Prompt-Based Learning

Cheikh Brahim El Vaigh*[1], Thomas Girault*
Cyrielle Mallart[1], Duc Hau Nguyen[1]
cheikh-brahim.el-vaigh@inria.fr,thomas@girault.fr
duc-hau.nguyen@irisa.fr,cyrielle.mallart@inria.fr
[1]Univ. Rennes, INRIA, CNRS, IRISA, France

ABSTRACT

In this paper, we present our participation to the MediaEval-2021 challenge on fake news detection about coronavirus related Tweets. It consists in three subtasks that can be seen as multi-labels classification problems we solved with transformer-based models. We show that each task can be solved independantly with mutiple monotasks models or jointly with an unique multitasks model. Moreover, we propose a prompt-based model that has been fine-tuned to generate classifications from a pre-trained model based on DistilGPT-2. Our experimental results show the multitask model to be the best to solve the three tasks.

1 INTRODUCTION

With the worldwide spread in the last few years of the Sars-Cov-2 virus, also known as Coronavirus, fear and concern has grown. While traditional media often relies on scientifically-vetted sources to bring information to concerned readers or viewers, social media is not subjected to this obligation to fact check. Therefore, a plethora of messages of various degrees of truthfulness has emerged across social media platforms, such as Tweeter. This network has been used as a soapbox for a multitude extreme political theories revolving around the Coronavirus epidemic, as well as for the debunking of said conspiracies, making it harder to untangle the conspiracy theories from the real facts.

This paper describes the systems [1] that we developed for MediaEval 2021 Fake News detection challenge. The main objective of the present task is to classify tweets according to whether they are relaying conspiracy theories, and what specific conspiracy is evoked. Several sub-tasks contribute to refining the task[7]: task 1 aims at deciding whether a tweet contributes to a conspiracy, mentions a coronavirus-related conspiracy, or is unrelated, task 2 aims at classifying the topic of a tweet, while task 3 combines the two previous labels into a multiclass problem, with both the relevance of the tweet and its subject to infer. For each sub-task, data sets were provided by the organizers of the challenge [8].

For each of the three aforementioned tasks, we propose a classification solution relying on the fine-tuning of transformer-based models e.g., [1, 3, 5, 6]. We also propose a prompt-based learning approach for the first task, relying on the DistilGPT-2 [9] pretrained model. Here, inferring the label of a tweet is treated as a text generation task, with the entirety of the tweet to classify as a prompt. Additionally, these three tasks being related, we propose a multitask approach that learns on all three sets of labels, and later allows for filtering the results for a specific subtask. This solution yields better results than separating all three subtasks.

2 APPROACH

We explain hereunder the different models we devised to address to detect conspiracy theories in tweets. We first describe the separated task learning framework in Sec. 2.1, than we introduce the Prompt-based model in Sec. 2.3. Finally, Sec. 2.1 is dedicated to the multi-task setup where all the three different task are performed at once.

2.1 Separated multilabel models for each task

The tasks can be solved with multi-labels model to classify a tweet as a vector of independent probabilities for each label thanks to a Sigmoid activation function. The labels of the task 2 are already well encoded as a binary matrix, whereas for the tasks 1 and 3, the original categorical labels have been converted to binary targets with one-hot-encoding.

For each task, a separate instance of several BERT-based models are fine-tuned. We tested bert-tiny [1], vaccinating-covid-tweets [2], a model based on BERTweet [6] fine-tuned on Covid related tweets, and toxic-bert [5]. The first model is used as a first baseline to compare the approaches, as it is a smaller version of the BERT model and requires a small amount of time to finetune. The BERTweet based model has already been trained on tweet-formatted documents, and may therefore learn on subtler aspects of tweets. Finally, toxic-bert has been chosen as it has been trained on toxic (hateful, obscene, threatening, etc.) tweets, and may therefore pick up on the fear-mongering language used by conspiracy theorists.

2.2 Multitask model

The multitask model uses the same backbone as the first approach with separated tasks. Each set of labels (for tasks 1, 2 and 3) are concatenated into a single set of labels. The idea behind this multitask approach is to learn one general model that can be used to perform the different tasks taking advantage of the relation between the tasks e.g., the existence of a conspiracy for task-1 or the existence of a particular conspiracy theory for task-3 (fine-grained version of task-1 and task-2). Thus by properly performing task-3, we expect the multitask model to be better in task-1 and task-2 as they are more general than task-3.

*These authors contributed equally to this work
[1]source code available online: https://github.com/CMallart/FakeNewsMediaeval2021

[2]https://huggingface.co/ans/vaccinating-covid-tweets

MediaEval'21, December 13-15 2021, Online

2.3 Prompt-based model

According to the prompt-based learning approach, samples are made into templates, containing the text of the tweet, a label for the tweet, and a binary classification of whether the tweet is related to its label, as in the following :

```
Tweet : Media succeeded in creating this Covid 19 hoax...
Label : Promotes/Supports Conspiracy
Classification : true
```

As shown in the previous template, we formulated this problem as a binary classification task. For each tweet, one template is created for each possible label.

The language model is then trained to output the final word, chosen in a list that consists of the words ["false", "unlinked", "unrelated", "true", "related", "linked"]. This word should be consistent with the previous prompt, which is the text and the label, and therefore correctly learn the type of tweets associated with each label. At inference time, three templates are again created for one tweet : the text of the tweet, followed by each of the possible three labels. The chosen label is the one where the model has the highest probability to output "true", "related" or "linked" as the following word.

This prompt-based model was implemented with the use of the OpenPrompt library [4], trained on 15 epochs. Due to lack of time and GPU resources, the prompt-based model has been only trained to task 1 but it could easily extended to multitask.

3 RESULTS AND ANALYSIS

Table 1 shows the results on task-1 of the different finetuned pretrained models, as well as the prompt-based approach. The best scores (MCC, macro-f1 and micro-f1) are obtained with the covid-tweet model which has been clearly optimized on a corpus adapted to our task. The toxic-bert reached comparable results but it appears that pretraining on toxic language is not really transferable to the conspiracy language. Surprisingly, smaller models such as bert-small and bert-tiny were able to achieve quite competitive results.

Prompt-based learning with DistilGPT-2 does not outperform traditional fine-tuning on the task 1. However, we would expect that this approach would benefits from expanding to multitask where the labels share semantic properties and the number of example per label is low.

As expected, the multitask approach outperforms the individual models on all three tasks, as displayed in Table 2. The detailed results for each label are given in table 3.

Table 1: Comparing different pre-trained models on task-1 in the separated tasks setup

Model	MCC	micro-F1	macro-F1
covid-tweets	0.58	0.73	0.71
toxic-bert	0.56	0.73	0.71
bert-small	0.55	0.73	0.71
bert-tiny	0.48	0.68	0.66
distilGPT-2 - prompt	0.45	0.65	0.63

Table 2: Comparing the multi-task and separated task models on dev set and the test set (official MCC), for the covid-tweets model

Approach	Task	MCC-test	MCC-dev	micro-F1	macro-F1
sep. tasks	1	0.60	0.58	0.73	0.71
	2	0.72	0.72	0.75	0.76
	3	0.66	0.68	0.94	0.67
multitask	1	**0.63**	0.70	0.81	0.79
	2	**0.73**	0.75	0.78	0.82
	3	**0.68**	0.77	0.95	0.75

Table 3: Detailed results on tasks 2 and 3 on the test set (MCC), for the multitask approach

Label	Task 2	Task 3
Suppressed cures	0.72	0.70
Behaviour and Mind Control	0.80	0.72
Antivax	0.68	0.62
Fake virus	0.65	0.64
Intentional Pandemic	0.50	0.51
Harmful Radiation/Influence	0.80	0.73
Population reduction	0.83	0.75
New World Order	0.86	0.80
Satanism	0.71	0.66
Official Run Score (official MCC)	0.73	0.68

4 DISCUSSION AND OUTLOOK

In this work, we experimented transformer-based models to detect conspiracy theories in tweets. The three tasks have been solved with multilabels classifiers relying on the pretrained models. We showed that it is better to train the models jointly on multiple tasks rather than independently. Meanwhile, the official MCC scores, while good, still show that there is large space for progress, especially for the task-1 which has only three labels. The other tasks are more challenging due to diversity of labels and the small size of the dataset. The idea of using the prompt-based model shows promising results for task-1, but due to lack of time and resources we would rather focus on the other experiments. As shown in [2], we also tried to generate fake training samples with GPT-2 but we were not able to use them due to the lack of annotations. In the future works, we plan to use a larger generative GPT-2 model for prompt-based training, apply the prompt model to the three tasks and try a multitask prompt based model, which will combine two of our promising approaches, to outperform the proposed multitask.

REFERENCES

[1] Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics. (2021). arXiv:cs.CL/2110.01518

[2] Vincent Claveau. 2020. Detecting Fake News in Tweets from Text and Propagation Graph: IRISA's Paritcipation to the FakeNews Task at MediaEval 2020. 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/paper63.pdf

[3] 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

[4] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. OpenPrompt: An Open-source Framework for Prompt-learning. *arXiv preprint arXiv:2111.01998* (2021).

[5] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify. (2020).

[6] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 9–14.

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

[8] 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 *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*. 21–25.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). arXiv:1910.01108 http://arxiv.org/abs/1910.01108