Detecting Fake News in Tweets from Text and Propagation Graph: IRISA's Participation to the FakeNews Task at MediaEval 2020

Vincent Claveau
CNRS, IRISA, Univ. Rennes, France
vincent.claveau@irisa.fr

ABSTRACT

This paper presents the participation of IRISA to the task of fake news detection from tweets, relying either on the text or on propagation information. For the text based detection, variants of BERT-based classification are proposed. In order to improve this standard approach, we investigate the interest of augmenting the dataset by creating tweets with fine-tuned generative models. For the graph based detection, we have proposed models characterizing the propagation of the news or the users' reputation.

1 INTRODUCTION AND RELATED WORK

This paper describes the systems that we developed for the text-based and structure-based MediaEval 2020 Fake News detection challenge. These two subtasks and the datasets are detailed in [10] and [12].

Text classification is a common NLP task [6]. Although simple machine learning approaches have shown promising results for fake news detection [8], the recent transformer-based architectures, such as BeRT [2], have set new standards. Several large pre-trained transformer models are now available; they are known to yield state-of-the-art results on many NLP tasks including text classification [16, *inter alia*]. We rely on one of these pre-trained models to build our systems. In order to improve this standard approach, we have investigated the interest of augmenting the dataset artificially by generating tweets with fine-tuned generative models (one for each class). These approaches and results are detailed in Sec. 2.

Similarly, classification of data represented as a graph, and in particular node classification, is not new but the recent trend is to use deep learning [5]. Yet, for the specific domain of fake news detection, other approaches are possible. In particular, it has been shown that the fake news are propagated differently (and faster) than legit news [15]. The use of node reputation and link-based analysis, as it is done in the detection of spam web pages from the Web graph (such as TrustRank [4], an adaptation of PageRank [1]) is another inspiration for our approaches. Our two approaches are further detailed in Sec. 3.

2 TEXT-BASED APPROACHES

2.1 Pre-processing

From the tweets still online[1], the text is extracted and pre-processed as follows. Emojis are transformed into texts [13]. URLs are changed to the fixed string `'URL'`. Twitter usernames are removed if they appear once, others are kept and the @ removed. The intuition is that some often cited users may be associated to a specific class. Hashtags are kept (with # removed), and decomposed when they contain a mix of capital and small letters (eg. `#CovidHoax` is changed in `CovidHoax Covid Hoax`).

2.2 Generating artificial examples

For this task we wanted to investigate the use of generative models in order to artificially augment and balance the datasets. Indeed, the performance of neural language models based on transformers [14] makes this task realistic. To do so, we use GPT2 (Generative Pre-Trained Transformers), a model built from stacked transformers (precisely, decoders) trained on a large corpus by auto-regression [11]. Three GPT2 models – one for each class – are fine-tuned (from the 355M-parameter pre-trained model) with the tweets of the dev set. The amount of tweets available is very small; we stopped the iterations when perplexity reached 0.5. The way this stopping criterion impacts the results would need further investigations, which were not possible due to the limited time of the challenge. For the generation, we randomly picked up tweets and kept the two first words to serve as bootstrap. The temperature, which controls the creativity of the model, was set at 0.7. Here again, we had no time to investigate the impact of this parameter. Approximately 20,000 tweets were generated for each class. Here are some tweets generated for the class '5G conspiracy':

```
Crude and unproductive! Turn off the 5G in your area and see if that helps. Covid19 is not funny. I hope that the Wuhan government puts an end to this immediately.
"Immigrants are the cause of 5G towers, they're the cause of the coronavirus outbreak, they're the covid-19 victims, the 5G towers are the weapon which will eradicate the world population, 5G lays the microchips for the virus, i read somewhere that the 5G was debuting prior to the introduction of the COVID-19 virus to negate some of the hype around COVID-19
```

2.3 Classification models

Our 4 classification variants are based on the RoBerta-large model [7]. It was preferred over other transformer-based representations because its tokenizer is expected to be more suited for the tweet writing specifics. We have tested models with different classification layers (SVM, logistic regression), with or without fine tuning, and with or without artificial examples. Finally, the submitted runs are the following ones:
model 1: tweet embedding from the Roberta model (not fine-tuned), and SVM (RGB kernel);
model 2: Roberta model with a linear classification layer, fine-tuned on the task (3 epochs);

[1] At retrieval time, respectively 227, 128 and 80 tweets were no longer available for the class 'non', '5G', 'other' in the dev set.

MediaEval'20, December 14-15 2020, Online

Table 1: Performance of the proposed systems for the text-based and graph-based detection; models are detailed in Sec. 2 and 3.

	cross validation results			official
model	MCC	micro-F1	macro-F1	MCC
model 1 (text)	0.4654	0.7460	0.5924	0.4680
model 2 (text)	0.5345	0.7945	0.6253	0.5571
model 3 (text)	-	-	-	0.4937
model 4 (text)	-	-	-	0.4888
reputation (graph)	0.4415	0.7274	0.5900	0.4093
propagation (graph)	0.3198	0.6051	0.4980	0.3036

model 3: same as model 2, with artificially generated examples (3 epochs);
model 4: same as model 3 (4 epochs).

2.4 Results of text-based detection

The results of our models are given in Tab. 1. When available, in addition to the official score on the test set, we provide Matthews correlation coefficient (MCC), micro-F1 (accuracy) and macro-F1 on the dev data (80% for training, 20% for validation). Note that due to the cost of the artificial example generation and the small amount of data, the GPT2 models are fine-tuned on all the available dev data; we do not have reliable results for models 3 and 4 (generated tweets added to the training set can be very similar to those in the validation set).

From the results, we see that fine-tuning the representation (model 2 vs. model 1) is beneficial. Unfortunately, the artificially generated tweets (model 3 and 4) do not yield the expected improvement. From the confusion matrices, one can see that the class 'other conspiracy' has the poorest results, with tweets being equally labeled as '5G', 'non' or 'other'.

3 GRAPH-BASED APPROACHES

For the second sub-task, we have proposed two models, based on two different sets of features. They are described in the following subsections, as well as the machine learning algorithms adopted and their results.

3.1 Modeling the user's reputation

This set of features aims at taking into account if one of the users posting or propagating the news has already be seen. Each user is indeed associated with a score for each possible label, computed from the numbers of training samples of each class it was associated with. We also take into account the scores of the neighbors of this user, their own neighbors, and so on... In practice, this is implemented with the PageRank algorithm [1] on the undirected graph with a dumping factor set to 0.8 (optimized by cross-validation). Finally, each sample ends up with one value for each class; these three scores are the features used by the classifier.

Several learning algorithms have been tested (logistic regression, random forests, SVM; as implemented in scikit learn [9]). The optimal settings for their hyper-parameters are grid-searched using 20% of the dev set as validation set. The weight of each sample is adapted according to the inverse of its class proportion ('balanced' strategy). With their optimal settings, the different learning algorithms finally show little differences. For this set of features, the submitted run was produced with a random forest (1,000 trees with a maximal depth set to 5, Out-of-Bag weights used in the prediction).

3.2 Modeling the propagation

This set of features is built by considering how the tweet is propagated (without considering the users' reputations). These features can be used even if every involved user has never been seen before and is not connected any known user. The features include (with n_0 the first user tweeting the piece of news): number of nodes in the propagation graph; total number of friends and followers (for all nodes implied), as well as the median, 25% percentile, 75% percentile of followers; number of followers and friends of n_0; difference between the number of followers and friends of n_0; maximal, minimal, average, median, 25% percentile, 75% percentile of retweet time; times to reach at least 100, 1,000, 10,000 followers and so on up to 200,000 followers. With this set of features, a SVM has been used with the following parameters: standardized features (removed mean and scaled to unit variance), RBF kernel, C=0.9, gamma automatically set with the 'scale' heuristics.

3.3 Results of graph-based detection

The results of the systems are given in Tab. 1. The cross-validation and official results are consistent; they both show the advantage of the reputation-based approach, especially when considering micro-F1. The difference between cross-validation and official test score may be explained by a lower amount of already seen nodes in the test set, compared to what was generated by cross-validation. A system exploiting all the proposed features (propagation + reputation) was also tested but obtained no statistical difference with the reputation only features.

For both models, the 'other conspiracy' class is again the most error-prone (proportionally), with an equal amount of the its tweets being classified in the three classes. Overall, for both feature sets, many errors are caused by confusion between the 5G and non 5G conspiracy tweets.

4 CONCLUSION AND FUTURE WORK

For the detection of fake news based on the text, we have adopted a state-of-the-art approach based on RoBerta. The scores obtained show that there exists a large margin for progress, especially when dealing with close classes (5G vs. other conspiracies). The idea of incorporating artificially generated examples did not result in better performance and still needs some work. First, we may find better ways to set the training and generation hyper-parameters. Secondly, we plan to investigate the use of generative model to expand the sample at inference time.

For the detection based on the structure, we have shown that simple approaches like reputation already offered promising results, even on small datasets with many unseen-before nodes. In addition to this type of approach, we want to explore more recent node representation techniques that make it possible to use deep learning, such as node2vec [3] or subsequent variants.

REFERENCES

[1] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web 7 (WWW7)*. Elsevier Science Publishers B. V., Brisbane, Australia, 107–117.

[2] 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: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[3] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 855–864. https://doi.org/10.1145/2939672.2939754

[4] Z. Gyngyi and H. Garcia-Molina. 2005. Link spam alliances. In *Proceedings of the 31st international conference on Very large data bases, VLDB*. Trondheim, Norway, 517–528.

[5] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation Learning on Graphs: Methods and Applications. *IEEE Computer Society Technical Committee on Data Engineering* (2017).

[6] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text Classification Algorithms: A Survey. *Information* (2019).

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019). arXiv:cs.CL/1907.11692

[8] Cédric Maigrot, Vincent Claveau, Ewa Kijak, and Ronan Sicre. 2016. MediaEval 2016: A multimodal system for the Verifying Multimedia Use task. In *MediaEval 2016: "Verfiying Multimedia Use" task*. Hilversum, Netherlands. https://doi.org/10.1145/1235

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[10] 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*.

[11] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* (2019).

[12] Daniel Thilo Schroeder, Konstantin Pogorelov, and Johannes Langguth. 2019. FACT: a Framework for Analysis and Capture of Twitter Graphs. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 134–141.

[13] Kevin Wurster Taehoon Kim. 2020. Emoji Python library. (2020). https://pypi.org/project/emoji/

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[15] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559 arXiv:https://science.sciencemag.org/content/359/6380/1146.full.pdf

[16] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.