Evaluating Standard Classifiers for Detecting COVID-19 Related Misinformation

Daniel Thilo Schroeder[23], Konstantin Pogorelov[1], Johannes Langguth[1]
[1]Simula Research Laboratory, Norway
[2]Simula Metropolitan Center for Digital Engineering, Norway
[3]Technical University of Berlin, Germany
daniels@simula.no,konstantin@simula.no,langguth@simula.no

ABSTRACT

This paper summarises the results created through participation in the task *FakeNews: Corona Virus and 5G Conspiracy* of the MediaEval Multimedia Evaluation Challenge 2020. The task consists of two parts intending to detect tweets and retweet cascades that emerged during the COVID-19 pandemic and causally connect the radiation of 5G networks with the virus. We applied several well-established neural networks and machine learning techniques for the first subtasks, namely, textual information classification. For the second task, the retweet cascades analysis, we rely on classifiers that work on established graph features, such as the clustering coefficient or graph diameter. Our results show a MCC-score of 0.148 or 0.162 for the NLP task and 0.02 for the structure task.

1 INTRODUCTION

The COVID-19 pandemic and the associated lockdown formed the basis for a multitude of false news and conspiracy myths. While a large amount of this content is limited to virtual spaces such as news portals or online social networks, some amount of particularly fast-moving content has immediate consequences in the real world. The *FakeNews*-task[9] at the MediaEval challenge 2020 targeted the classification of tweets and retweet cascades of such a so-called *Digital Wildfires.* More precisely, the claim that COVID-19 is causally related to the radiation emitted by 5G towers, which began to spread rapidly during February 2020 and later, in early April, culminated in attacks on telecommunication staff and arson against 5G masts. This naturally raises the question whether Tweets promoting such dangerous misinformation can be detected automatically, either by analyzing tweet contents or the network between Twitter accounts that spread such misinformation, and the cascades created by the spread.

While the analysis of tweets is an established problem, the examination of retweet cascades is the subject of ongoing research. Retweet cascades can be obtained from Twitter without much effort[11], that their quality can be improved[4]. Previous work indicates that the false content spreads faster, deeper, and wider than true content[12] and that accounts that support similar ideas tend to form clusters in the Twitter graph which are sometimes referred to as *echo chambers* [2, 7].

While natural language processing has advanced rapidly in the last three years [3, 5], such NLP techniques rely on massive models. Similarly, retweet cascades can be analyzed using graph neural networks, but they too tend to be computationally expensive. Our goal in this paper is to assess the performance of established methods that are both simple and computationally cheap for both subtasks.

2 SUBTASK 1: NLP-BASED DETECTION

For the NLP-based detection subtask we decided to evaluate an entry-level complexity of the tweet classification based only on plain text content. Thus we implemented the simple out-of-box multi-class classifier based on the standard implementation of *Support Vector Machine* (SVM) from the *scikit-learn* framework [8]. During our experiments, we observed that the execution of the SVM classifier for both single- and multi-class cases on the raw text data extracted from the corresponding tweeter posts resulted in worse-than-random performance. A quick look into the downloaded training data showed that the text blocks of most tweets are heavily contaminated by non-informative character noise, emoticons, and various URLs. To overcome this problem, we implemented a multi-stage text cleaning procedure that performs the following steps on each text block: remove all control characters; remove all hyperlinks; glue all sequences of non-numerical and non-letter characters into one symbol; remove all hashtags and user references; remove all non-punctuation symbols and remove all short words. As a result, single- and multi-class SVN classifiers performed more efficiently, giving better than random classification results during the cross-validation on the provided development set. After a series of experiments, we selected the two best-performing configurations of SVM-classifiers (see Table 1 for the classifier configurations).

Table 1: NLP-based classifier parameters

Run	Classifier	Kernel	Ngram	Vertorizer
NLP Single-Class	SVC	linear	(1,6)	TF-IDF
NLP Multi-Class	SVC	linear	(1,12)	TF-IDF

3 SUBTASK 2: GRAPH-BASED DETECTION

Our graph-based detection approach is relatively straightforward and assumes that individuals spreading false messages tend to be organized in so-called homophile networks. Homophile networks are more strongly connected internally than externally, which relates to echo chambers and related concepts which can lead to resistance to ideas coming from outside the homophile network [13]. However, because Twitter does not allow access to true retweet cascades (i.e. for each retweet list the specific account whose existing retweet caused the new retweet) but instead returns just a list of retweeters for a particular tweet, this challenge offers the subgraphs of Twitters' follower network that were induced by these retweeters. Induced subgraphs mirror the neighborhoods of the tweet authors and we intend to treat them as such, meaning that we rather want to study the relationship between individuals rather than the content distribution over time[10]. Therefore, we define a mapping from each retweet cascade to a point in $\mathbb{R}^5$ by calculating the following measures: *(1)* the average local clustering coefficient, which for each vertex quantifies how close its neighbors are to being a clique, *(2)* the main component to components ratio, which after running a connected component analysis divides the number of nodes in all other components by the number of nodes in the main component, *(3)* the diameter, which is the longest shortest path from the original tweet author to any retweeter vertex in the main component, *(4)* the number of vertices, and *(5)* the number of edges. Vosoughi and Roy[12] have shown that false news spreads faster, deeper, and wider than true news. We argue that *(3, 4, 5)* is a measure of depth and breadth, while *(1, 2)* reflects the aspect of homophile networks. In this sense *(2)* appears to be an interesting measure since it reflects a multi-medial spreading on the one hand, because the existence of multiple components indicates that the spreading did not occur along with Twitter's follower network, and a wider and deeper spreading on the other hand. We then used the well-known classifiers Naive Bayes and Random Forest to train two models using the retweet cascade representation described above.

4 RESULTS

In total we have submitted two runs for the NLP-based Detection and one run for Graph-based Detection subtasks. The official evaluation results for the submitted runs, which are shown in Table 2, depict the expected picture of the evaluation performance. Here we obtained a relatively low MCC score since we intentionally selected the most simple-to-implement approaches with relatively simple custom additions to the initial data preprocessing. The NLP-based Detection showed almost equal performance for single- and multi-class classification runs: an MCC score of 0.148 and 0.162 respectively. On the one hand, these scores depict the complexity of the NLP tasks, especially for social-media-scrapped data that can contain both straightforward target statements as well as sarcastic and humoristic posts that are hard to recognized even for an experienced human evaluator. On the other hand, the fact that our improved yet simple SVM-classifier is able to perform above the random baseline (MCC of 0.0) gives clear evidence that the NLP-task presented in this challenge has high research and educational potential for both experienced and beginner researchers.

Furthermore, the evaluation score for the Graph-based Detection subtask of 0.02 is just above the random classifier performance. This can be seen as a confirmation of the high complexity of tasks related to graph structure analysis and graph property matching. We conclude that simple approaches are not suitable for the structure-oriented classification tasks, especially under the conditions of the limited size of the available training dataset.

Table 2: Evaluation results on the test dataset

Submitted Run	Official Multi-class-MCC Score
NLP Single-Class	0.148
NLP Multi-Class	0.162
Structure Multi-Class	0.020

5 DISCUSSION AND OUTLOOK

We successfully implemented an out-of-the-box approach with the additional text data pre-processing to NLP classification problem to prove that NLP-tasks can be tackled by both experienced and inexperienced audiences. However, the achieved multi-class MCC score of 0.162 gives a clear indication that more advanced models are required to address the NPL-analysis problems for social-media-generated data. In future we plan to implement more complex NLP models, e.g. BERT word embedding [3] or similar techniques. Our approach to structure-based misinformation detection is based solely on examining an authors' environment and makes no reference to the time aspect at all. We argue that this strategy is a valid approach to classify conspiracy and non-conspiracy tweets, while it seems like the classification between Digital Wildfires and non-fast-spreading conspiracies should aim for a model that takes time, speed, or acceleration into account. Furthermore, assuming that certain network motifs[6] appear in more than one retweet cascade seem to be appropriate since recent research[1] has shown that disinformation campaigns partly organize the distribution of misinformation which may lead a set of Twitter accounts that are used multiples time with hardcoded retweeting in between each other. The last two points suggest that it makes sense to embed the graphs in a space where time exists. However, we leave this part open for further research.

ACKNOWLEDGMENTS

This work was funded by the Norwegian SAMRISK-2 project "UMOD" (#272019). The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[2] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication* 64, 2 (2014), 317–332.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. 2012. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce.* 623–638.

[5] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504* (2019).

[6] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.

[7] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2020. Shared Partisanship Dramatically Increases Social Tie Formation in a Twitter Field Experiment. (2020).

[8] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[9] 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*.

[10] Daniel Thilo Schroeder, Pedro Lind, Konstantin Pogorelov, and Johannes Langguth. 2020. A Framework for Interaction-based Propagation Analysis in Online Social Networks. (12 2020).

[11] 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.

[12] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[13] L Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. 2012. Competition among memes in a world with limited attention. *Scientific reports* 2 (03 2012), 335. https://doi.org/10.1038/srep00335