FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020

Konstantin Pogorelov[1],Daniel Thilo Schroeder[13],Luk Burchard[4],
Johannes Moe[12],Stefan Brenner[5],Petra Filkukova[1],Johannes Langguth[1]
[1]Simula Research Laboratory, Norway[2]University of Oslo, Norway
[3]Simula Metropolitan Center for Digital Engineering, Norway
[4]Technical University of Berlin, Germany[5]Stuttgart Media University, Germany
{konstantin,daniels,langguth,petrafilkukova}@simula.no,l.burchard@campus.tu-berlin.de,\sb288@hdm-stuttgart.de,
arenor.moe@gmail.com

ABSTRACT

The FakeNews: Corona Virus and 5G Conspiracy task, running for the first time as part of MediaEval 2020, focuses on the classification of tweet texts and retweet cascades for the detection of fast-spreading misinformation, and therefore provides a low-threshold introduction to natural language processing and graph analysis. This paper describes the task, including use case and motivation, challenges, the dataset with ground truth, the required participant runs, and the evaluation metrics.

1 INTRODUCTION

Digital wildfires, i.e., fast-spreading inaccurate, counterfactual, or intentionally misleading and information that can quickly permeate public consciousness and have severe real-world implications, are among the top global risks in the 21st century [10]. While misinformation is widespread on the internet, only a very small portion of it leads to harmful harmful acts in the real world. In2020. the COVID-19 pandemic has severely affected people worldwide and consequently dominated world news for months. Thus, it is no surprise that it has also been the topic of a massive amount of misinformation, which was most likely amplified by the fact that many details about the virus were unknown at the start of the pandemic. We are particularly interested in detecting content associated with a Digital Wildfire that relates COVID-19 to 5G wireless technology and led to arson and attacks on telecommunications workers. Despite the emphasis on COVID-19 and 5G, we further differentiate between content that does not contain misinformation and content attributed to other misinformation. Our task offers two subtasks: The first subtask includes text-based tweets classification, while the second targets the classification of retweet cascades [11].

In contrast to text-only classification challenges, e.g., [1, 7, 13], our dataset also contains retweet cascades, allowing us to consider diffusion as a characteristic shown to be valuable for the spread of misinformation [20]. The final goal is the inclusion of various field experts aiming for efficient multi-modal approaches. Furthermore, we ask for evaluation of different approaches utilizing both as little and as much training data as possible and evaluating the approaches with respect to real-world imbalanced datasets [2].

There are already many methods for automatic news analysis and fake content detection in the social media and news analysis field, e.g. [3, 5, 12, 16] that cover a wide range of approaches, including knowledge graphs, diffusion models, and natural language processing. These methods typically rely on labeled data. Consequently, several such datasets have been published in recent years [4, 6, 8, 14, 15, 17, 19, 21]. However, to our best knowledge, there is no existing dataset that emphasizes Digital Wildfires and takes retweet cascades into account.

The task is intended to be of interest to researchers in the areas of online news, social media, multimedia analysis, multimedia information retrieval, natural language processing, and meaning understanding and situational awareness.

2 DATASET DETAILS

Our dataset's creation can roughly be divided into four steps. First, we used Twitters'search API between January 17, 2020 and May 15, 2020 to collect a large number of statuses (i.e. *tweets*, *retweets*, *quotes*, and *replies*) including key-words related to the COVID-19 pandemic. Moreover, we filtered for those that mention 5G in any conceivable spelling such as 5G, 5g, or #5g. Second, we restored as much of the Twitter threads as possible using our custom framework [18]. The result is a graph of tweets, retweets, and quotes that does not only consist of statuses containing the obvious combination of keywords but provides more subtle content like *"All this to declare martial law huh? Lol or do you wanna put fear in us so we can run and get them vaccines to make us suseptible for these damn 5g tower radiations? Lol either way, the government is to NOT be trusted and are up to something."*. Some threads containing these statuses have their origin long before the COVID-19 pandemic. Nevertheless, we decided to include this data, too, since it contains context that led to our Digital Wildfires' emergence. In the third step, based on the number of statuses obtained in steps one and two, we started the manual labeling. Therefore, we randomly selected a subset of $10k$ tweets with their corresponding retweets. The annotation process has been performed by a team of researchers, postdocs, Ph.Ds, and master students. Each team member received an part of the subsets and these data were then annotated manually. Most of the easy-to-annotate statuses were assessed and classified by one annotator, but when assigning a class was not obvious, the tweet was discussed with the entire group until consensus was reached. While the text dataset was prepared via manual labelling, extracting the retweet cascades requires an additional step. A cascades root is always a labeled tweet while all other nodes correspond to retweets. We again made use of Twitter's API to fetch these retweets and the underlying social network that connects users via follower

MediaEval'20, 14-15 December 2020, Online

relationships. Unfortunately, Twitter limits the number of available retweets which narrows the cascade size to one hundred. Since each tweet and retweet contains a timestamp, one can track the temporal diffusion. However, Twitter does not provide the true retweet path, thus leaving it to the challenge participants to reconstruct it. We use three classes to label tweets and retweet cascades: **The 5G-Corona Conspiracy** class corresponds to all tweets that claim or insinuate some deep or obvious connection between COVID-19 and 5G, such as the idea that 5G weakens the immune system and thus caused the current Corona-virus pandemic, or that there is no pandemic and the COVID-19 victims were actually harmed by radiation emitted by 5G network towers. The crucial requirement is the claimed existence of some causal link. **The Other Conspiracy** class corresponds to all tweets that spread conspiracy theories other than the ones discussed above. This includes ideas about an intentional release of the virus, forced or harmful vaccinations, or the virus being a hoax. **The Non-Conspiracy** class corresponds to all tweets not belonging to the previous two classes and includes those discussing COVID-19 pandemic itself, claiming that 5G is not proven to be absolutely safe or even can be harmful without linking it to COVID-19, as well as claiming that authorities are pushing for the installation of 5G while the Publicis distracted by COVID-19. In addition, tweets pointing out the existence of conspiracy theories or mocking them fall into this class since they do not spread the conspiracy theories by inciting people to believe in them.

2.1 Dataset Contents

The development and test datasets consist of $6,458$ tweets and $2,327$ retweet graphs, and $3,230$ tweets and $1,165$ retweet graphs respectively, stored in two folders each: tweets and graphs. Both datasets are heavily unbalanced in terms of the number of samples per class, reflecting the distribution of tweet topics and people's opinions. To comply with the Twitter data publication policy, we provide only tweet IDs, but not the tweet text itself. An additional tweet content download script is provided to obtain the tweets from their ids via the corresponding Twitter API using a user-supplied API access keys. Retweet cascades are stored individually in a separate folder with three files. The *edges.txt* file contains a directed edge list source-node-ID to target-node-ID. The *plot.png* file contains a plot of the cascade. The *nodes.csv* contains an assignment from the node ID to the following properties: *id* - an anonymized node ID which remains the same for all graphs in the dataset of all categories; *time* - the time difference in seconds from each retweet to the original tweet. The original tweet always has a difference of 0 seconds to itself; *friends* - the next greater power of two of the follower count from the user profile of the respective user; *followers* - the next greater power of two of the friend count from the user profile of the respective user.

3 EVALUATION METRICS AND SUBTASKS

The officially reported metric used for evaluating the multi-class classification performance is the multi-class generalization of the Matthews correlation coefficient (MCC) [9]. In case of equal metric values, we use the timestamp of the official run submission to rank the teams. For the evaluation, the participants must submit one run for both subtasks defined below. Additionally, they optionally can submit four more runs for any of the described subtasks, i.e., participants can submit up to ten runs in total.

Text-Based Misinformation Detection Subtask: In this subtask, the participants are asked to perform classification of the tweets based on the tweet text contents and other tweet-relevant multimedia and meta-information can be obtained from Twitter or the Internet. The subtask requires one mandatory and four optional runs to be submitted. The required run implements a pure NLP classification of tweets based only on tweet text content without using any additional sources of data. Optional runs gradually extend the amount and types of allowed additional information implementing classification based on tweet text analysis in combination with visual information (images and/or videos) extracted from the original tweet and classification using any automatically scraped data from any external sources.

Structure-Based Misinformation Detection Subtask: In this subtask, the participants are asked to perform a classification of tweet graphs based on the tweet retweet graph, and additional retweet-tree-related information was obtained from Twitter. The subtask requires one mandatory and four optional runs to be submitted. The required run implements a pure tweet classification based only on the retweet graph structure only, without using any additional data. Optional runs gradually extend the amount and types of allowed additional information implementing classification based on a full set of retweet graph description, retweeting nodes' properties, and using any automatically scraped data from any external sources.

Thus, the participants are allowed to use only information that can be extracted from the provided tweets (including metadata) and retweet cascades for generating the first and second run for both subtasks. In contrast, for other runs everything is allowed, both from the data collection method perspective and the sources of information used. However, manual annotation of tweets or any externally scraped data is not allowed in any run.

4 DISCUSSION AND OUTLOOK

The task itself can be seen as very atypical and challenging due to a fairly limited amount of information available to support the tweet classification process. This reflects the real-world conditions in which online social media analysis systems are deployed. Thus, this task is a practical attempt to make a step towards building a usable multi-modal social network analysis system that is able to combine isolated data source properties with inter-source relations. Due to the importance of the use case, we hope to motivate researchers from different research fields to present their approaches, thereby performing research that can help society to fight against malicious manipulations of social networks and threats to society in general. We hope that the FakeNews task can help to raise awareness of the topic, but also provide an interesting and meaningful use case to researchers interested in this application.

ACKNOWLEDGMENTS

This work was funded by the Norwegian SAMRISK-2 project "UMOD" (#272019). It 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] 2018. Toxic Comment Classification Challenge - Identify and classify toxic online comments. (2018). https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/

[2] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter* 6, 1 (2004), 1–6.

[3] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. https://doi.org/10.1145/3394486.3403092. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 492–502. https://doi.org/10.1145/3394486.3403092

[4] Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 853–862.

[5] Dylan de Beer and Machdel Matthee. 2020. Approaches to Identify Fake News: A Systematic Literature Review. In *Integrated Science in Digital Age 2020*, Tatiana Antipova (Ed.). Springer International Publishing, Cham, 13–22.

[6] Sameer Dhoju, Md Main Uddin Rony, Muhammad Ashad Kabir, and Naeemul Hassan. 2019. Differences in health news from reliable and unreliable media. In *Companion Proceedings of The 2019 World Wide Web Conference*. 981–987.

[7] Quan Do. 2019. Jigsaw Unintended Bias in Toxicity Classification. (2019).

[8] Amira Ghenai and Yelena Mejova. 2018. Fake cures: user-centric modeling of health misinformation in social media. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–20.

[9] Jan Gorodkin. 2004. Comparing two K-category assignments by a K-category correlation coefficient. *Computational biology and chemistry* 28, 5-6 (2004), 367–374.

[10] Lee Howell. 2013. Digital Wildfires in a Hyperconnected World. https://bit.ly/2GiEF4f. (2013).

[11] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. 2012. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2335–2338.

[12] Thai Le, Suhang Wang, and Dongwon Lee. 2020. MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models. *arXiv preprint arXiv:2009.01048* (2020).

[13] Akshay Mungekar, Nikita Parab, Prateek Nima, and Sanchit Pereira. 2019. Quora insincere question classification. *National College of Ireland* (2019).

[14] Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2515–2519.

[15] Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854* (2019).

[16] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104* (2017).

[17] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. 2019. Fa-kes: A fake news dataset around the syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 573–582.

[18] 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.

[19] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286* 8 (2018).

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

[21] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).