FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021

Konstantin Pogorelov[1],Daniel Thilo Schroeder[13],Stefan Brenner[5],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}@simula.no,sb288@hdm-stuttgart.de

ABSTRACT

The FakeNews: Corona Virus and Conspiracies Multimedia Analysis task, running for the second time as part of MediaEval 2021, focuses on the classification of tweet texts aiming detection of fast-spreading misinformation. Task of this year extends the number of target conspiracy theories and introduces new challenges in terms of analysis complexity of the imbalanced dataset. 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

During the development of the COVID-crisis, a lot of new COVID-related conspiracy theories have arise. Despite efforts of the major social networks, mass-spread fake facts, irrational theories and news-like posts are widely presented in the online media sources. Rumors and other fast-spreading inaccurate, counterfactual, or intentionally misleading information can quickly permeate public consciousness and have severe real-world implications. Public attention to the problem have already allowed content moderation and partial limitation of freedom of speech in order to prevent manipulation of COVID-related public opinion. Thus, fake news and intentional missinformation are still among the top global risks in the 21st century [6]. Consequentially, we are particularly interested in detecting content associated with the fake news and COVID-related missinformation. We further differentiate between content that does not contain misinformation and content attributed to other misinformation. Our task offers three subtasks, all require text-based tweets classification.

Similar to text-only classification challenges, e.g., [1, 4, 7], we expect to see NLP approaches for tweet text analysis, but we aim wider set of conspiracy theories and different-level detection methodologies. Furthermore, we ask for evaluation of different approaches with respect to real-world imbalanced datasets [3].

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 datasets creation can roughly be divided into four steps. First, We used Twitters' search API between January 17, 2020 and Jun 30, 2021 to collect a large number of *tweets* that include keywords related to the COVID-19 pandemic. Second, we started [8] the manual labeling of randomly selected subset of approximately $2k$ tweets. The annotation process has been performed by a team of researchers, postdocs, PhDs, and master students. Each tweet was annotated by at least two annotators. Disagreed annotations war resolved by a third experienced annotator. In cases when assigning a class was not obvious, the tweet was discussed with the entire group until consensus was reached.

We use three classes to label tweets:

Promotes/Supports Conspiracy class contains all tweets that promotes, supports, claim, insinuate some connection between COVID-19 and various conspiracies, such as, for example, the idea that 5G weakens the immune system and thus caused the current corona-virus pandemic; that there is no pandemic and the COVID-19 victims were actually harmed by radiation emitted by 5G network towers; ideas about an intentional release of the virus, forced or harmful vaccinations, vaccine contains microchips, or the virus being a hoax, etc. The crucial requirement is the claimed existence of some causal link.

Discusses Conspiracy class contains all tweets that just mentioning the existing various conspiracies connected to COVID-19, or negating such a connection in clearly negative or sarcastic manner.

Non-Conspiracy class contains all tweets not belonging to the previous two classes. Note that this also includes tweets that discuss COVID-19 pandemic itself.

We use the following nine categories that corresponds to the most popular conspiracy theories: **Suppressed cures, Behaviour and Mind Control, Antivax, Fake virus, Intentional Pandemic, Harmful Radiation or Influence, Population reduction, New World Order, and Satanism.**

The development and test datasets consist of 1,554 and 266 tweets respectively. Both datasets are heavily unbalanced in terms of the number of samples per class, reflecting the distribution of tweet topics and people's opinions. The development dataset was divided into pre-flight and primary development sets. Pre-flight development set was provided earlier than primary and thus used to perform the initial approach selection and further as a validation set. To comply with the Twitter data publication policy, no data was publicly shared during the active challenge phase. Thus, all the registered participants are, in fact, become a closed group of researchers working together on one topic. To become a member of the research team all the registered participants are obliged to sign an additional strict NDA agreement. Within this research, we provide only tweet text content without any linking to the user accounts or original tweets. The full-text datasets have not been

MediaEval'21, December 13-15 2021, Online

made publicly available and they are sent to the members of the research team via the direct emails.

After the challenge, the annotated datasets containing only tweet IDs, but not the tweet text itself will be made publicly available. These publicly available datasets will be shuffled and supplied by the additional content to prevent linking to the full-text datasets was used during the challenge by the researcher team. An additional tweet content download script will be provided to obtain the tweets from their ids via the corresponding Twitter API using a user-supplied API access keys.

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, Rk-statistic) [5]. This metric provides an efficient and reliable comparison for multi-class classifiers for both balanced and unbalanced datasets.

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 at least one run for at least one subtask defined below. Additionally, the participants optionally can submit four more runs for any of the described subtasks, i.e., participants can submit up to 15 runs in total.

Text-Based Misinformation Detection: In this subtask, the participants receive a dataset consisting of tweet text blocks in English related to COVID-19 and various conspiracy theories. The participants are encouraged to build a multi-class classifier that can flag whether a tweet promotes/supports or discusses at least one (or many) of the conspiracy theories. In the case if the particular tweet promotes/supports one conspiracy theory and just discusses another, the result of the detection for the particular tweet is expected to be equal to "stronger" class: promote/support in the given sample.

Text-Based Conspiracy Theories Recognition: In this subtask, the participants receive a dataset consisting of tweet text blocks in English related to COVID-19 and various conspiracy theories. The main goal of this subtask is to build a detector that can detect whether a text in any form mentions or refers to any of the predefined conspiracy topics.

Text-Based Combined Misinformation and Conspiracies Detection: In this subtask, the participants receive a dataset consisting of tweet text blocks in English related to COVID-19 and various conspiracy theories. The goal of this subtask is to build a complex multi-labelling multi-class detector that for each topic from a list of predefined conspiracy topics can predict whether a tweet promotes/supports or just discusses that particular topic.

All the subtask, in which the team has decided to participate, requires one mandatory and four optional runs to be submitted. The required mandatory 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 by implementing classification based on tweet text analysis in combination with pre-trained models and classification using any automatically scraped data from any external sources. Manual annotation of tweets or any externally scraped data is not allowed in any run.

In the submitted runs participants are allowed to use an additional *Cannot Determine* class. This additional class represents cases, when the output of the classifier is not reliable. This additional class is important for evaluation of multi-class classifiers. The effect of using *Cannot Determine* class is described in the related literature [2]. In-short, marking a sample that classifier cannot reliable classify as an unknown class affects the resulting classification performance less negatively than marking the sample with a wrong class label, exactly as it expected to be implemented in a real-world classification tasks.

With respect to the subtasks evaluation, the following methodology is used. **Text-Based Misinformation Detection** subtask is evaluated with Rk-statistic directly. **Text-Based Conspiracy Theories Recognition** and **Text-Based Combined Misinformation and Conspiracies Detection** subtasks are evaluated with the two-steps evaluation procedure. First, evaluation of each conspiracy theory individually and independently is performed using Rk-statistic. Then all the computed Rk-statistic values across all the conspiracy theories are averaged and the resulting averaged value is used to compare results of different teams. Finally, results in each conspiracy theory group are evaluated independently, but this evaluation is auxiliary and do not affect the final teams ranking.

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 Research Council under contracts #272019 and #303404 and 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. We also acknowledge support from Michael Kreil in the collection of Twitter data.

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] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one* 12, 6 (2017), e0177678.

[3] 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.

[4] Quan Do. 2019. Jigsaw Unintended Bias in Toxicity Classification. (2019).
[5] Jan Gorodkin. 2004. Comparing two K-category assignments by a K-category correlation coefficient. *Computational biology and chemistry* 28, 5-6 (2004), 367–374.
[6] Lee Howell. 2013. Digital Wildfires in a Hyperconnected World. https://bit.ly/2GiEF4f. (2013).
[7] Akshay Mungekar, Nikita Parab, Prateek Nima, and Sanchit Pereira. 2019. Quora insincere question classification. *National College of Ireland* (2019).
[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.