Combining Tweets and Connections Graph for FakeNews Detection at MediaEval 2022

Konstantin Pogorelov[1,*,†], Daniel Thilo Schroeder[1,†], Stefan Brenner[3,†], Asep Maulana[1,†] and Johannes Langguth[1,2,†]

[1]*Simula Research Laboratory, Norway*
[2]*University of Bergen, Norway*
[3]*Stuttgart Media University, Germany*

Abstract

The FakeNews Detection task at MediaEval 2022, running for the third time as part of the challenge, focuses on the detection of misinformation tweets and their spreaders. Like in the 2021 task, conspiracy theories related to COVID-19 in nine different categories have to be detected, along with the authors stance towards them. However, the size of the dataset has approximately doubled. Furthermore, we also provide a large interaction graph along with vertex features derived from the same Twitter dataset in which misinformation spreaders should be classified. As a final subtask, participants are asked to combine text and graph information to refine their classifications. This paper describes the tasks, including use case and motivation, challenges, the dataset with ground truth, the required participant runs, and the evaluation metrics.

1. Introduction

During the course of the COVID-19 pandemic a large amount of misinformation of various kinds was observed in online and offline media of all kinds. A particularly noteworthy example of this misinformation are conspiracy theories related to the origin, nature, and treatment of the virus. Despite the efforts of most major social networks, irrational and or harmful conspiracy theories spread widely in many online media, and the spread of such content can have severe real-world implications. Thus, our aim is to study new ways of detecting such content, as well as the stance of the author towards the content. We are especially interested in messages that propose multiple overlapping conspiracy theories.

However, conspiracy content can be difficult to detect by pure text analysis, since many such ideas are communicated via hidden or implied meaning, codes, or intentional misspellings such as *plANdemic* instead of *pandemic*. Thus, in order to improve detection accuracy, we suggest to study the connections between tweet authors, as well as meta-information about them. Thus, our task offers three subtasks, with the first requiring text-based tweet classification, the second node classification, and the third a combination thereof.

Similar to text-only classification challenges, e.g., [1, 2, 3], we expect to see NLP approaches for the 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

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online
*Corresponding author.
†These authors contributed equally.
konstantin@simula.no (K. Pogorelov); daniels@simula.no (D. T. Schroeder); bionescu@imag.pub.ro (S. Brenner); asep@simula.no (A. Maulana); langguth@simula.no (J. Langguth)

real-world imbalanced datasets [4]. The second subtask requires graph analysis, using graph neural networks or similar methods, and consequently the third subtask requires both.

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 semantic understanding.

2. Dataset Details

The dataset was created in a multi-stage process, starting with the collection of a set of tweets related to the COVID-19 pandemic from Twitter between January 17, 2020 and Jun 30, 2021. We used the Twitter search API via our custom distributed Twitter scrapping framework called *FACT* [5] and targeted COVID-19 keywords. Since conspiracy tweets are not particularly frequent, we use a list of keywords related to conspiracy theories and perform a text search. During the COVID-19 pandemic we observed misinformation trends and developed the list. We then removed tweets that contain hyperlinks. This was done because using the links could distract from the goal of the challenge, i.e. natural language understanding. For the remaining tweets, we attempted to resolve the self-reported location of the tweet authors. We make use of a system to resolve locations from previous work [6]. From the remaining set, we selected $3,389$ tweets and performed the manual labeling. The selection was done in a way that ensures that a constant proportion of the tweets was selected from every day in the dataset, in order to ensure an even distribution and to account for the fact that the daily number of COVID-19 related tweets was much higher in Spring 2020 than during the later stages of the pandemic.

The annotation process was 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 in nine categories 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 tweet text datasets consist of $1,913$ and $1,476$ 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 whole development dataset is used in the first and third subtasks. The test dataset is devided between the first and third subtasks resulting in two subsets containing 830 and 646 tweets respectively. Tweet texts were shared with the

registered participants who were obliged to sign an additional non disclosure agreement.

For the first subtask, we provide only tweet text content without any linking to the user accounts or original tweets objects. On the other hand, for the second task we provide a graph with $1,679,011$ vertices and $268,694,698$ edges, along with $1,913$ and 830 vertex labels for the development and test set respectively. In addition, it contains more details about the users. The account creation date, number of favourites, followers, friends, and statuses, self-reported location resolved via the Google geolocation API, and the verification status. On the other hand, userID, name and description have been anonymized. For the latter two, the length of the string is given instead. Thus, it should not be possible to identify individual users with simple methods such as Google searches. For the third task, we also provide a matching between annotated tweets and annonymized users.

After the challenge, the entire tweet dataset will be made available, but only tweet IDs, labels, and graphs will be shared publicly. Tweet texts can be obtained by contacting the authors. A paper describing the dataset is in preparation [7].

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) [8] which is suited for multi-class classifiers for both balanced and unbalanced datasets. Ties are broken by submission time. 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 each subtasks, for a maximum of 15 runs in total.

Text-Based 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. This task is identical to a task posed in last year's challenge, but it uses a larger development and test datasets.

Graph-Based Conspiracy Source Detection: In this subtask, the participants are given an undirected graph derived from social network data where the vertices are users and the edges represent connections between them. Each vertex has a set of attributes, including location, number of followers, as well as some texts posted by that user. Some users are labeled as misinformation posters, based on manually annotated tweets, and some are labeled as non-misinformation posters. This subtask asks participants to classify the other users in the graph, based on their connection to the labeled users as well as their attributes. Scoring will be based on correctly classifying users/vertices in the graph that have manually generated hidden labels.

Graph and Text-Based Conspiracy Detection: This subtask combines the data of both previous subtasks with the aim of improving the text-based classification. For each text to be evaluated, the vertex corresponding to the author is specified in the graph. The goal of this subtask is the same as that of Subtask 1, but participants can make full use of the graph data and vertex attributes. This subtask will use the same development and a different test set from that of the first subtask.

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. The effect of using the *Cannot Determine* class is described in the related literature [9].

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 in two-steps. 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 step is auxiliary and do not affect the final ranking.

4. Discussion and Outlook

The task is substantially more challenging than the 2021 edition [10], with last years task being contained in the first subtask. It resumes the use of graphs as a tool to improve detection accuracy from the 2020 challenge [11] but by using a large connected graph instead of individual spreading graphs, we open the way for trying out new network based approaches.

5. Acknowledgements

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] Q. Do, Jigsaw unintended bias in toxicity classification (2019).
[2] Toxic comment classification challenge - identify and classify toxic online comments, 2018. URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/.
[3] A. Mungekar, N. Parab, P. Nima, S. Pereira, Quora insincere question classification, National College of Ireland (2019).
[4] N. V. Chawla, N. Japkowicz, A. Kotcz, Special issue on learning from imbalanced data sets, ACM SIGKDD explorations newsletter 6 (2004) 1–6.
[5] D. T. Schroeder, K. Pogorelov, J. Langguth, Fact: a framework for analysis and capture of twitter graphs, in: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2019, pp. 134–141.
[6] J. Langguth, P. Filkuková, S. Brenner, D. T. Schroeder, K. Pogorelov, Covid-19 and 5g conspiracy theories: long term observation of a digital wildfire, International Journal of Data Science and Analytics (2022) 1–18.
[7] J. Langguth, D. T. Schroeder, P. Filkuková, S. Brenner, J. Philips, K. Pogorelov, Coco: An annotated twitter dataset of covid-19 conspiracy theories (2022).
[8] J. Gorodkin, Comparing two k-category assignments by a k-category correlation coefficient, Computational biology and chemistry 28 (2004) 367–374.
[9] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using matthews correlation coefficient metric, PloS one 12 (2017) e0177678.
[10] K. Pogorelov, D. T. Schroeder, S. Brenner, J. Langguth, Fakenews: Corona virus and conspiracies multimedia analysis task at mediaeval 2021, in: Multimedia Benchmark Workshop, 2021, p. 67.
[11] K. Pogorelov, D. T. Schroeder, L. Burchard, J. Moe, S. Brenner, P. Filkukova, J. Langguth, Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020., in: MediaEval, 2020.