Katholieke
Universiteit
Leuven

**Faculty of
Science**

ANALYTICAL REPORT

Collecting & Analysing Big Data for Social Sciences [S0K17A]
Coordinator: prof. Dr. Cecil Meeusen
M.Sc. Statistics and Data Science

(GROUP – 12)

Prabhjyoth MATTUMMAL (r0861984)
Jakub CIEROCKI (r0867514)
Aleksandra KOTOWICZ (r0878531)

Academic Year 2021–2022

Contents

1 Critical Analysis

1.1 Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal

1.1.1 Overview

The article examines the relationship between COVID-19 vaccination rates, vaccine hesitancy, and online misinformation about vaccines.

The researchers used several data sources to collect tweets and information about them: CoVaxxy dataset available on GitHub and Python Carmen library [1, 2]. By using CoVaxxy dataset, the researchers were able to collect 55 million tweets from January 4 to March 25 using Twitter API and a list of vaccine-related keywords. Carmen library on the other hand, which helps with determining Twitter API message location, allowed scientists to geolocate approximately 1.67 million users from 50 US states and approximately 1.15 million users from over 1300 counties, which then allowed to get around 11 million shared tweets.

To find disinformation, researchers searched for mentions and links to articles from websites identified by a politically neutral third party as unreliable. In addition, they measured the widespread presence of vaccine misinformation in each region.

Their dependent variables were vaccination uptake rates at the state level (reported as the number of daily vaccinations administered in each state for the week from March 19 to 25, 2021 and data collected from the CDC), and vaccine hesitancy at the state and county levels (based on the Facebook Symptom Survey by Delphi Group at Carnegie Mellon University).

The models used in the research were implemented in Stata 16. Based on the available data, the scientists performed multivariate regression models adjusted for six potential confounders: percentage of population below the poverty line, percentage of 65+, percentage of inhabitants in each racial and ethnic group, rural-city continuum code, number of COVID -19 deaths per thousand and per cent of Republican votes (in units of 10 per cent). Other covariates, including religiosity, unemployment rate and population density, were also taken into account. Researchers also performed various sensitivity analyses, including different specifications of the disinformation variable and logged versions of disinformation, and multiple regression models predicting vaccine rate and vaccination hesitancy using weighted least squares regression, as both dependent variables were normally distributed. Scientists compared two autoregressive models to investigate Granger's causal relationship between vaccine hesitancy and misinformation. The first model included daily rates of vaccine hesitancy x at time t in geographic region r (state or county), and the second model added daily rates of misinformation per account to an exogenous variable y.

The research found a negative correlation between online misinformation and vaccination uptake rates and a relationship between misinformation and vaccine hesitancy rates based on survey data. Moreover, the relationship between vaccines and misinformation remains significant regarding demographic, socioeconomic and political factors. The Granger causality analysis showed a directional relationship between misinformation and vaccine hesitancy.

1.1.2 Remarks

Online misinformation regarding matters like COVID-19 can be harmful to society. It not only may cause increased vaccine hesitancy but as well vaccine refusal. It all may be a barrier affecting achieving herd immunity [3]. This is why the topic that researchers in the article "Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal" studied is really

relevant, and the research can be as well later used to try to increase social awareness on the topic of misinformation about vaccines.

When it comes to the study, its design seems to be standard and suitable. The researchers used many sources, like tweets from Twitter (which are a common way to collect public opinion views due to their accessibility), publicly available data from the Centers for Disease Control and Prevention (CDC), and Facebook Symptom Surveys data provided by the Delphi Group at Carnegie Mellon University. It allowed them to look for different views of people (based on social media data) as well as connect the social opinions with vaccination data. Moreover, the methods that were applied are as well common and used in researching similar issues [4, 5, 6].

The used data sources in the research are well-chosen, but they as well have some limitations. Take a look at the first dataset, CoVaxxy, which is a collection of English-language Twitter posts about COVID-19 vaccines. The researchers working on this collection used the standard way to collect data using Twitter API and keywords since the Twitter platform allows its users to do it. Nevertheless, the dataset provided by CoVaxxy may have some limitations, as it is essential to remember that Twitter users are only part of the population and may not be a representative sample of it. The same goes for their opinions, as they may not represent everyone's views [7, 8]. Moreover, to collect a dataset, the scientists chose only tweets in English, which can exclude some minority dialect speakers, multilingual speakers or people speaking in other languages [9, 8]. Furthermore, there might have been a problem that while collecting such a large dataset, the users did not have any chance to opt-out of it, which may provoke some ethical questions about anonymity [8]. As it can be seen, the used dataset provided by researchers from CoVaxxy collection may have some limitations, but as of now, collecting that much data on such a large scale from Twitter is tough to perform. Another used data was Facebook Symptom Surveys provided by the Delphi Group at Carnegie Mellon University. The survey has more than 22 million responses, which seems to be a reasonable number, and the researchers had a good idea to include the data from it [10, 11]. Even though it is a valuable dataset, it has some limitations. It has the same limitation as CoVaxxy collection as Facebook users, who took part in the study, may not be a representative sample of the population. Additionally, as the authors of the article mention, even though there were no missing survey data regarding vaccine hesitation at the state level, observations were missing at the county level, which may limit the study. The researchers also considered the limitation connected with the fact that Facebook survey data are available only when the number of respondents is at least 100 and also used that limit dealing with Twitter accounts geolocated in each county. Additionally, they decided to use a more extended time window to measure vaccine hesitancy (January 4th - March 25th, 2021) as they assumed that it might affect uptake rates, which was a reasonable decision.

It is worth noticing that the whole study would still be possible today, as all the used sources are still available. Moreover, the data collection would be as well possible, for example, from tweets, as the Twitter API is still available to users, and there has been no limitations added to it since the time the study was conducted. Because of that, such an analysis is still possible and reproducible. In addition, the scientists who worked on the research are open about the different steps being taken during research and provide the manual and necessary codes on GitHub to process data and analyse it [12].

To ensure that findings are robust to alternative variables and model specifications, the researchers conducted a set of sensitivity analyses. It was a good decision to check if the data or models had any problems.

The project is interesting. Even though the research may have some limitations, for example, regarding data, it is useful and can be used later to raise awareness on the issue of misinformation concerning vaccines.

1.2 Automatic detection of emerging COVID-19 conspiracy theories in social media and the news

1.2.1 Overview

Researchers developed a scraper to collect data published from Reddit forums (subreddits) and 4Chan threads related to the pandemic. Subreddits and threads were evaluated for relevance by three independent evaluators and were selected only by consensus.

The interlocking computational methods described in this study facilitate the discovery of a series of important features of the (i) narrative frameworks that bolster conspiracy theories and their constituent rumours circulating on and across social media, and (ii) the interaction between social media and the news.

The dataset contains almost a month's worth data from 4Chan and GDELT. Communities from the social media corpus were explored within the subset of news media between the same dates using Relative Coverage Scores. Then, the cross-correlation of the ratio of coverage scores for different fixed communities to a random community has been taken.

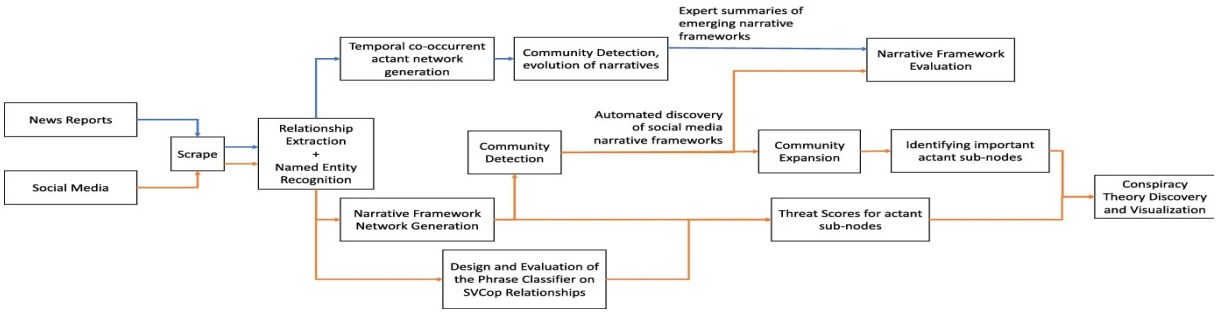


Figure 1: Pipeline of processing the data

Inspired by narrative theory, this study includes crawling around social media sites and news reports and, through the application of automated machine-learning methods, discover the underlying narrative frameworks supporting the generation of rumors and conspiracy theories.

For news reports, this research relied on the GDELT project, an open source platform that traverses web news (in addition to printing and broadcasting) from around the world gdeltproject.org. This dynamic news corpus report included a primary search for conspiracy theory. The corpus was then filtered to include only English articles (GDELT built-in features) from US news sources. These articles were then cleaned and staged for this study's pipeline to extract sentence-level relationships between key actors. It also extracted almost 60 relationships from each report, close to 50 filtered news reports per day, and 324510 relationships.

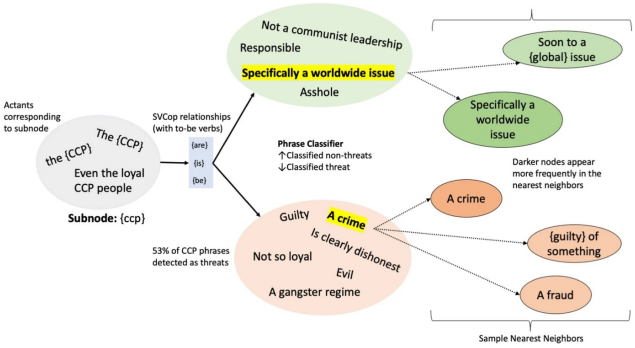
The latent structure of the social media networks also provides features which enable the identification of key people, places and things in conspiracies and conspiracy theories, and the detection of threat elements in these narratives.

The relationship between social media debates and conspiracy theories in media coverage changed during the research period. When the outbreak of coronavirus appeared to be confined to the city of Wuhan in central China from mid-to-late January and the United States was barely threatened, news media reporting the conspiracy theory reported the outbreak of coronavirus. As the outbreak continued until March 2020, conspiracy theoretic coverage gradually converged on the broader outbreak coverage. In mid-April, the focus shifted to reporting conspiracy theories discussed on social media, such as this certain research corpus.

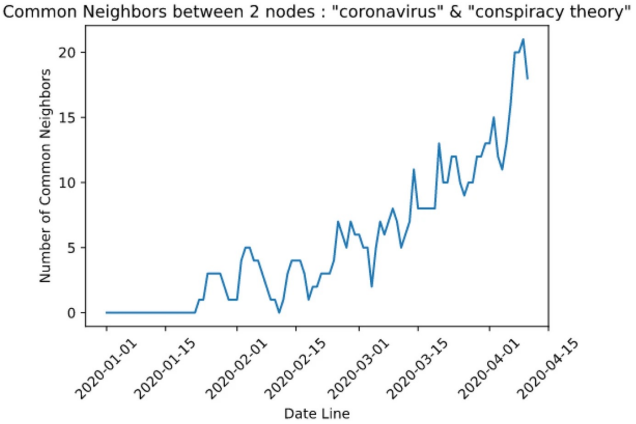
1.2.2 Remarks

Conspiracy theories thrive in low-confidence, low-trust environments. As a result, considering the lack of scientific consensus on the virus's transmission and containment, as well as the pandemic's long-term societal and economic repercussions, it's hardly unexpected that ones tied to the COVID-19 epidemic are flourishing. The virus is activated by the 5G telecommunication network, the pandemic is a hoax perpetrated by a global cabal, the virus is a bio-weapon released deliberately by the Chinese, or Bill Gates is using it as cover to launch a broad vaccination program to facilitate a global surveillance regime, according to stories currently circulating in US-focused social media forums.

The methods used by the researchers here are a refinement of those developed for an earlier study of conspiracies and conspiracy theories. They estimate narrative networks that represent the underlying structure of conspiracy theories in a large social media corpus (4Chan, Reddit) where they are most likely to originate, and the corresponding reporting about them in the news (GDELT). This approach allowed the researchers to analyse the interplay between the two corpora and to track the time-correlation and pervasive flow of information from one corpus to the other.



(a) Red clouds are classified as threats; green clouds are non-threats. The kNN classifier reasonably clusters phrases that are syntactically different but semantically similar.



(b) News reports: Across all 101 segments of 5-day intervals.

The threat classifier in this study is trained on the relationships extracted from social media posts [13, 14]. This approach definitely has certain limitations, such as data acquisition, narrative frame estimation, threat labelling, validation of extracted narrative graphs, and support for real-time analysis using pipelines [15].

Data extracted from social media sources tends to be very noisy, including heavy spam, irrelevant, non-topic conversations, and numerous links and images dotted with meaningful textual data. Even with cleanup, many text extracts are compromised by spelling, grammar and punctuation errors, and poor syntax. These issues are primarily addressed by the NLP module, but the extraction of entities and relationships in the social media corpus is less accurate than in the news corpus. Unlike news articles, which tend to be archived well, social media posts, especially on sites like 4Chan, are volatile because users often delete or hide posts. As a result, re-crawling the website can create significantly different target records. Furthermore, we believe that the semi-supervised approach used in the study for the threat detection would require less human effort if more accurate semantic embeddings were available.

Given the turbulent debate on social media surrounding the COVID-19 pandemic, the corpus seems to have several competing conspiracy theories. Therefore, researchers are expected to

find many communities throughout the network. Some communities are isolated from others, and others have a limited number of shared subnodes. Also, this network is expected to have a hierarchical structure.

As with many studies on social media, there is no clear basis for evaluating or validating it. This issue is especially evident in the context of folklore genres such as rumours, legends, and conspiracy theories, as there is no standard version of a particular story. Indeed, folklore is always a dynamically negotiated process and is based on the concept of variation, so it is not clear what the ground truth of any of these stories is. To address this issue, the narrative frameworks that emerge from social media could be considered and can be compared them to those that emerge from the news media. Verification of the results confirmed that the social media graph was more accurate than the news media graph.

There seems to be a fairly lively interaction between "Twittersphere" and the rest of the social media landscape, especially Facebook. Many tweets refer to discussions on social media, especially Facebook. However, this phenomenon couldn't be explained analytically because access to Facebook data for research purposes is restricted. Future work will include tweets linking to rumours and other conspiracy theories in the targeted areas of social media. As part of this integration, this study can also include considerations for the credibility of various Twitter nodes and the amplifying role that "bots" can play in spreading these stories [16].

Currently, this particular pipeline only works with English-language content. However, the modularity of the pipeline allows one to include language-specific NLP tools for analysing languages such as Italian and Spanish. Both areas are strongly pandemic-influenced and may have their own rumours and conspiracy theories, which may be an interesting follow-up study.

1.3 Synergistic discussion

Both research papers tackle a very similar subject, which is connected with online misinformation and conspiracy theories revolving around COVID-19. The topic is important as there is a critical need to develop computational algorithms for distinguishing false news from true news. Such techniques could be used to help organisations fact-check, discover and prevent the spread of false information [13].

The articles employ several forms of data to demonstrate various points of view. Twitter posts, administrative data from the CDC, and a Facebook survey are used in the first article, while data from forums and news is used in the second article. Additionally, both studies use the English language data, which is not surprising given how widely the language is spoken around the world. The projects also employ a variety of strategies, demonstrating that diverse approaches may be utilised to handle comparable problems. The first research paper focuses more on regression techniques, while the second article is more inclined towards the clustering approach.

The papers also show quite a few approaches to the ethical issue. Privacy and ethical problems are almost completely ignored in the first article. In the second article, the researchers observe that the news or mass media frequently appear to chase social media conversations about misinformation and random conspiracy theories created by strangers, and then proceed jumping to incorrect conclusions before making any clarifications, which is potentially unethical.

References

[1] CoVaxxy - GitHub. https://github.com/osome-iu/CoVaxxy. Accessed: 2022-06-18.

[2] Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. Carmen: A Twitter Geolocation System with Applications to Public Health.

[3] Philip Gerretsen, Julia Kim, Lena Quilty, Samantha Wells, Eric E. Brown, Branka Agic, Bruce G. Pollock, and Ariel Graff-Guerrero. Vaccine Hesitancy Is a Barrier to Achieving Equitable Herd Immunity Among Racial Minorities. *Frontiers in Medicine*, 8, 2021.

[4] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(10):201199.

[5] Anasse Bari, Matthias Heymann, Ryan J. Cohen, Robin Zhao, Levente Szabo, Shailesh Apas Vasandani, Aashish Khubchandani, Madeline DiLorenzo, and Megan Coffee. Exploring Coronavirus Disease 2019 Vaccine Hesitancy on Twitter Using Sentiment Analysis and Natural Language Processing Algorithms. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 74(Suppl_3):e4–e9, May 2022.

[6] Srikanth Umakanthan and Sam Lawrence. Predictors of COVID-19 vaccine hesitancy in Germany: a cross-sectional, population-based study. *Postgraduate Medical Journal*, February 2022.

[7] Stefan Wojcik and Adam Hughes. Sizing Up Twitter Users, April 2019.

[8] Matthew R. DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:992–999, May 2021.

[9] David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[10] Francesco Pierri, Brea L. Perry, Matthew R. DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports*, 12(1):5966, April 2022.

[11] Delphi Epidata API. https://github.com/cmu-delphi/delphi-epidata. Accessed: 2022-06-18.

[12] Reproducibility code for "Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal". https://github.com/osome-iu/CoVaxxy-Misinfo. Accessed: 2022-06-18.

[13] Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R. Tangherlini, and Vwani Roychowdhury. Conspiracy in the time of Corona: Automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*, 3(2):279–317, 2020.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[15] Julien Kervizic. Real-time data pipelines-complexities; considerations, Dec 2020.

[16] Emilio Ferrara. #covid-19 on twitter: Bots, conspiracies, and social media activism. *CoRR*, abs/2004.09531, 2020.

[17] The GDELT project. https://www.gdeltproject.org/.

[18] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537, November 2011.

[19] Wendy Lehnert Yale University, Wendy Lehnert, Yale University, and Other MetricsView Article Metrics. Narrative text summarization: Proceedings of the first aaai conference on artificial intelligence, August 1980.