Vaccine Misinformation – A Twitter Pandemic

A study of the spread of COVID-19 vaccine misinformation on a social-media platform

Ola Klingberg
Graduate student at
University of Colorado Boulder
ola.klingberg@colorado.edu

ABSTRACT

In this project, I investigate tweets containing misinformation about the COVID-19 vaccines. The project has two parts. In the first part, I use topic modeling, a natural language processing technique, to extract the main topics making up this misinformation, and compare them to the topics in vaccine-related tweets that don't contain misinformation. In the second part, I use network analysis to examine how and to what extent the main spreaders of vaccine misinformation are connected to each other through retweeting. I use a custom metric to compare how much each account contributes to the spread of the information, and I show that just 15 accounts are responsible for over 50% of the total spread. These insights are valuable when planning public-information campaigns and other measures to counter vaccine misinformation.

KEYWORDS

COVID-19
vaccine misinformation
vaccine hesitancy
Twitter
tweet

INTRODUCTION

Vaccines have been a tremendous boon to human health. In most parts of the world, children today don't have to fear polio or diphtheria, and nowhere does anybody have to worry about smallpox, which has been entirely eradicated, thanks to comprehensive vaccination campaigns. But this progress is not necessarily irreversible (except for the case of smallpox). When vaccination rates decline in a region, the communicable diseases of yore often make a comeback, taking life and causing preventable suffering.

As most vaccines are not 100% effective, part of the protection they offer is that they contribute to herd immunity; each individual in a vaccinated population is not guaranteed to be absolutely immune, but since the general level of immunity is so high, the disease is much less likely to spread through the population. Those with less-than-perfect immunity are protected by the immunity of those around them. Herd immunity also protects those who are unable to get vaccinated, such as children below a certain age (different for different vaccines) and people with compromised immune systems. A person's decision not to get vaccinated, or not have their children vaccinated, does therefore not just affect the health of the unvaccinated individuals, but contributes to a public health hazard affecting everyone around them.

While COVID-19 had a fairly low fatality rate, it has still been estimated that COVID-19 vaccines spared between 14.4 to 19.8 million lives worldwide during just the first year after it began being administered [1]. We don't know when the next pandemic will hit, and we also don't know what fatality rate it will have. But it's likely that vaccines will be an essential part of the response.

Given such high stakes, it's important to look at the reasons why people choose not to get vaccinated against preventable diseases. A main reason is misconceptions about vaccines. Examples include the debunked theory that the MMR vaccine (against measles, mumps, and rubella) can cause autism, or that the COVID-19 vaccines altered the recipient's DNA. While some misconceptions about vaccines have been around for at least several decades, their reach has increased with the advent of social media, where misinformation can spread fast and wide.

In this context it's important to note that not all vaccine hesitancy is based on misinformation. While the commonly used vaccines today have very low rates of serious side effects, they all carry some risk of some side effects. Someone may very well hesitate to take a vaccine based on a correct understanding of the possible side effects.

A note about terminology: the sets of tweets in this study come from 2020 and 2021, when the social-media platform was still called Twitter. I will therefore refer to it as Twitter, and to the posts as tweets, except when referring to the company today, when I will call it by its current name: X.

RELATED WORK

Hayawi et al. [2] present what they believe is the first dataset for detecting COVID-19 vaccine-related misinformation. The dataset consists of two subsets:

- 15M unlabeled vaccine-related tweets, from December 1, 2020 till July 31, 2021.
- 15k vaccine-related tweets, manually labeled based on whether or not they contain misinformation, from the same time period.

Along with this dataset, they present a machine-learning algorithm to classify COVID-19 misinformation. With a BERT (Bidirectional Encoder Representations from Transformers) classifier trained on the 15k manually labeled examples, they reach an impressive 0.98 test-set F1-score.

Cotfas et al.[3] explore the degree of vaccine hesitancy expressed in a total of over 5.3 million English-language tweets from two one-month periods following the discovery of the Delta and Omicron variants of COVID-19 (end of 2020 and summer of 2021, respectively). They use Latent Dirichlet Allocation (LDA) to extract topics from these sets, and then compare the differences in topics between these two periods. They make available a set of 5.7k tweets manually labeled based on their stance towards vaccinations: negative, neutral, or positive.

Both Hayawi et al. [2] and Cotfas et al. [3] are focused entirely on the texts of the tweets. Neither study looks into the tweeters behind the tweets and how the tweeters might be interconnected. A common action on X/Twitter is the retweeting, i.e. the forwarding of someone else's tweet to your own followers. In the second part of this study I look at who is retweeting whom, and build a network of the most impactful vaccine misinformers, where impact is defined using a custom metric.

METHODOLOGY

Topic Modeling

The labeled vaccine-misinformation dataset presented by Hayawi at a. [2]. was divided into two subsets based on the misinformation label:

- 3,825 tweets (35.7%) containing vaccine misinformation
- 6,884 tweets (64.3%) not containing vaccine misinformation

Latent Dirichlet Allocation (LDA) was used for topic modeling on these two subsets separately. Unigrams (i.e. tokens made up of a single word, e.g. "therapy"), and bigrams (i.e. tokens made up of two words, e.g. "gene therapy"), were considered; n-grams of higher order were not considered. The stopword list from the NLTK library was used to remove common words. Some custom stopwords were removed as well, as they seemed so universal to the thoughts expressed in this dataset as to not offer any information that could be used to separate topics, e.g. the words "vaccine" and "COVID". Tokens that occurred in less than 0.5% or more than 50% of the tweets were excluded.

The sets of topics generated with LDA were evaluated using C_v and C_npmi scores for coherence and Jaccard Diversity for diversity. Hyperparameters and number of topics were adjusted until sets of interpretable topics were achieved. Getting a fairly small number of topics (9 for misinformation tweets; 5 for non-misinformation) was prioritized over getting the best possible metrics. The automatically generated topic labels were lightly edited, mainly by omitting some words that seemed too general to provide any information.

Network Analysis

The network analysis is focused on the vaccine-misinformation tweets, the tweeters behind them, and the accounts that retweeted them.

Standard centrality measures for networks, such as eigenvector centrality or pagerank, show how influential the members of the network are compared to each other. But that is not exactly the focus here. Rather, what is of interest here is what portion of the total spread of vaccine misinformation in the dataset each member is responsible for. To measure this, a fairly simple custom metric was used, which we will call misinformation impact. Misinformation impact estimates the total number of

instances of a particular member reaching someone else with a misinformation tweet. The metric is made up of two parts: direct impact, and retweeted impact.

Direct impact

The number of misinformation tweets the tweeter has posted, multiplied by the tweeter's follower count.

Retweeted impact

The sum of a retweeter's follower count, for each retweet.

Example

A has 10,000 followers and posts two misinformation tweets.

- Direct impact: 10,000 * 2 = 20,000.

One of the tweets is retweeted by B, who has 3,000 followers. The other tweet is also retweeted by B, and also by C, who has 800 followers.

- Retweeted impact:
  - Tweet 1: 3,000
  - Tweet 2: 3,000 + 800 = 3,800
  - Total Retweeted Impact: 3,000 + 3,800 = 6,800
- Total misinformation impact: 20,000 (direct impact) + 6,800 (retweeted impact) = 28,600.

While this number approximates the total number of times that A has reached someone else with a misinformation tweet, all those tweets will of course not actually be read.

All tweeters in the dataset were ranked in descending order based on their misinformation impact. A network was built out of the 100 most impactful misinformers, along with any account that retweeted one of their misinformation tweets, as long as the retweeter had a follower count of at least 10,000.

EVALUATION

Topic Modeling

Based on the LDA analysis, the following topics have been found in the misinformation subset of the labeled vaccine-misinformation dataset:

- Topic 1—*Experimental and untested*: experimental, untested, virus, rushed;
- Topic 2—*Experimental therapy*: experimental, therapy, gene, long, term, mRNA;
- Topic 3—*Experimental gene therapy*: experimental, gene, therapy, Pfizer, child;
- Topic 4—*Experimental gene therapy*: gene, therapy, experimental, mRNA, Pfizer, government;
- Topic 5—*Depopulation and Bill Gates*: depopulation, Gates, Bill, poison, agenda;
- Topic 6—*Experimental gene therapy*: experimental, gene, therapy, Pfizer, research;
- Topic 7—*Experimental therapy* and *Freedom and force*: experimental, person, free, force;
- Topic 8—*Experimental gene therapy*: experimental, Pfizer, gene, therapy, death, vaccinate, Moderna;
- Topic 9—*Bioweapon and Depopulation*: bioweapon, therapy, gene, depopulation, immunity, herd;

Based on the LDA analysis, the following topics have been found in the non-misinformation subset of the labeled vaccine-misinformation dataset:

- Topic 1—*Got the shot*: got, first, dose, second, today, shot;
- Topic 2—*Vaccinated and grateful*: vaccinated, grateful, thank, vaccination;
- Topic 3—*Worry and excitement*: worry, excited, parent, appointment;
- Topic 4—*Worry*: worry, need, ever, vaccination, drink
- Topic 5—*Worry* and *Mask*: worry, ever, need, wear mask, social

The lists of topics in the non-misinformation and the misinformation subsets are quite distinct from each other, without any apparent overlap. Within the lists, on the other hand, there is a lot of overlap among the topics, even between topics that appear distinct from each other in the Intertopic Distance Map (next page). It seems that the idea of the vaccines being experimental plays such a large part in the misinformation conversation that it is not just included, but actually dominates, most of the topics. It would probably make sense to merge topics 2, 3, 4, and 8 into one topic: *Experimental gene therapy*: experimental, gene, mRNA, Pfizer, death.

The non-misinfo topics show perfect separation in the Intertopic Distance Map, but the list of topic labels nevertheless shows "worry" being the most salient term in three out of five topics.

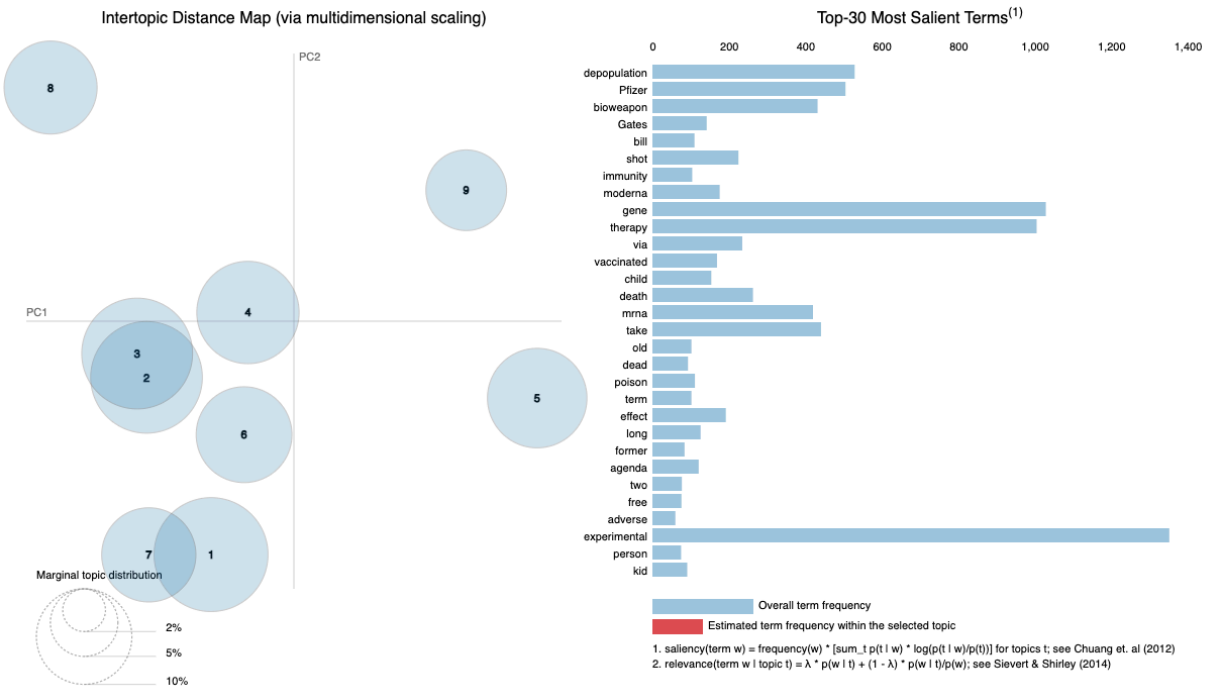


Figure 1. Latent Dirichlet Allocation topics and salient words in the misinformation subset of the labeled vaccine-misinformation dataset. [Link to interactive chart](#).

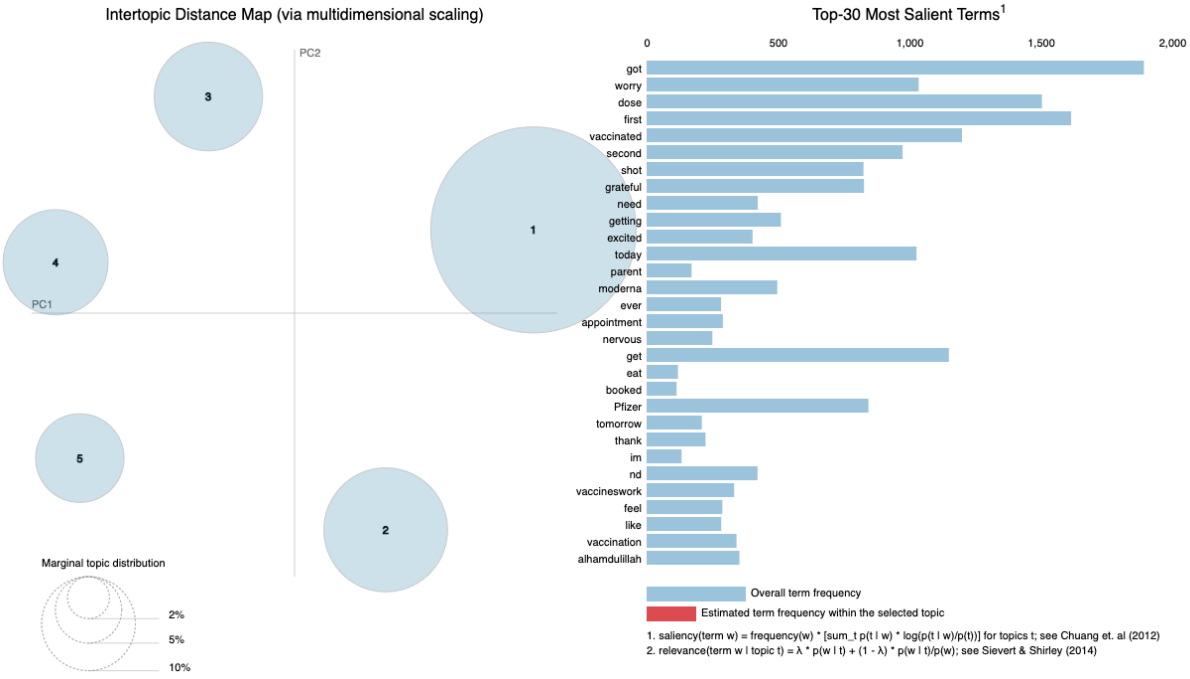
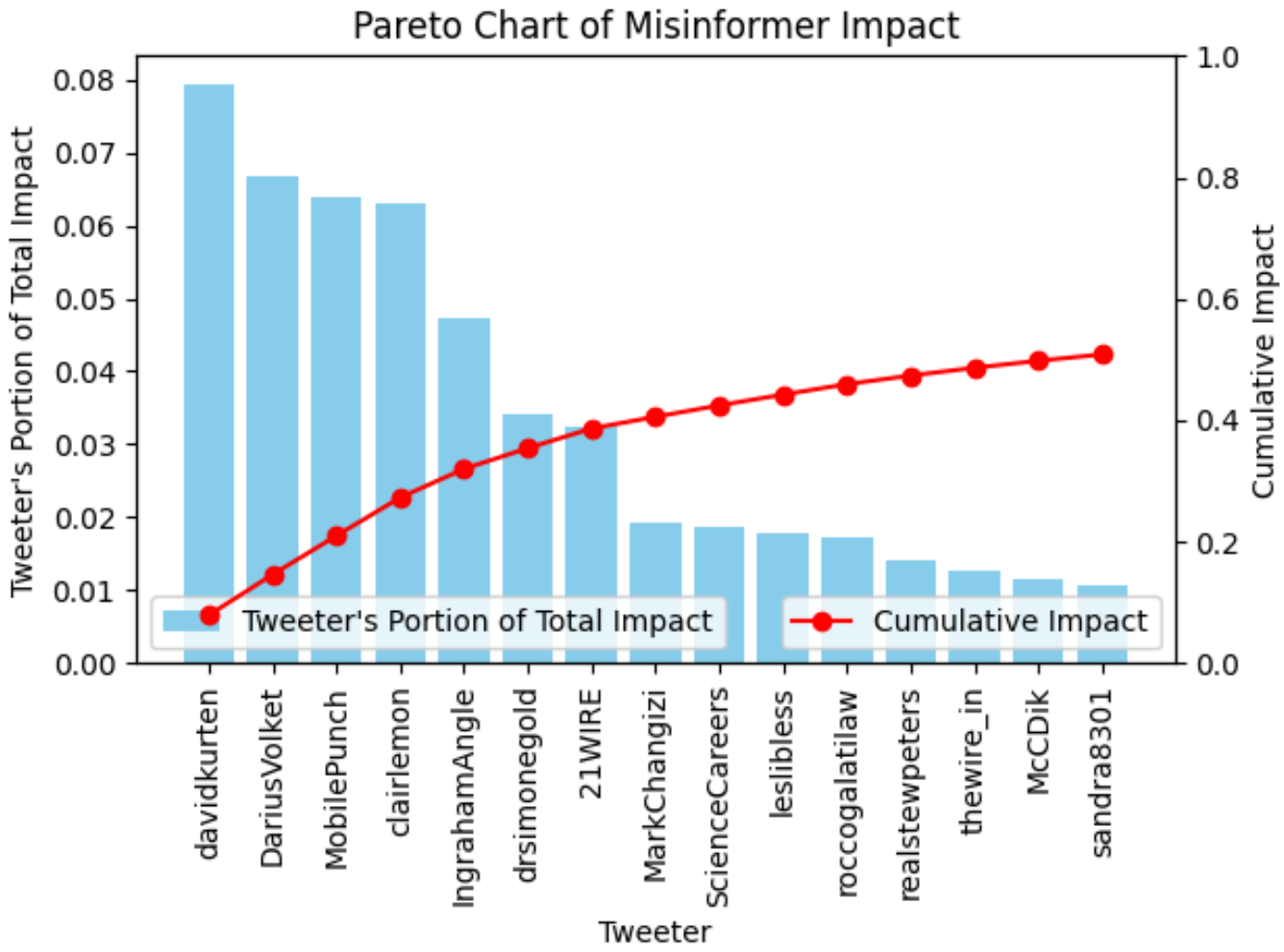


Figure 2. Latent Dirichlet Allocation topics and salient words in the non-misinformation subset of the labeled vaccine-misinformation dataset. [Link to interactive chart](#).

Network Analysis

The Pareto chart below shows the 15 tweeters with the largest misinformation impact, as defined above. As can be seen in the chart, these 15 tweeters together account for just over 50% of the total misinformation impact in the dataset.

A word of caution: in the first calculation of misinformation impact, the most impactful misinformer turned out to be Forbes, which is the account of Forbes Magazine. This was due to the size of their follower count, 20M, and three tweets with identical texts, which Hayawi et al. [2] had labeled as misinformation. After inspecting these tweets, I consider them mislabeled. They read: “Here are the differences between the Pfizer/BioNTech and Moderna Covid-19 vaccines and gene therapy”, and link to a Forbes Magazine article, with the headline “Covid-19 mRNA Vaccines Are Not ‘Gene Therapy,’ As Some Are Claiming.” The belief that mRNA vaccines constitute a form of ‘gene therapy’ is one of the common misconceptions about the COVID-19 vaccines. But these tweets don’t spread that misconception; rather, they refute it. Because of this, I removed Forbes from the analysis, but there is a further warning here: Once automated methods have yielded a small number of tweeters or tweets as being of special interest, it’s important to have these inspected, or re-inspected, manually by domain experts.



Vaccine Misinformation – A Twitter Pandemic

The network graph on the next page shows the most impactful misinformers, along with their most important retweeters. Accounts that posted original misinformation tweets are shown in red (whether or not they also retweeted others' tweets). Accounts that didn't post any of the original misinformation tweets in the dataset, but only retweeted others', are shown in gray. The size of the nodes are proportional to the square root of the number of followers the account has. Isolated nodes, of which there are four in the graph, represent accounts that are among the 15 most impactful misinformers, but who had no retweets that qualified for inclusion in this graph (i.e. any retweeters they may have had less than 10,000 followers).

A couple of observations that can immediately be drawn from the graph:

- Even though the account of JackPosobiec didn't post any of the original vaccine misinformation tweets in the dataset, it still contributed significantly to the spread of misinformation by retweeting tweets from an account with much fewer followers.
- A handful of accounts were retweeted often, giving them a reach far beyond their own set of followers.

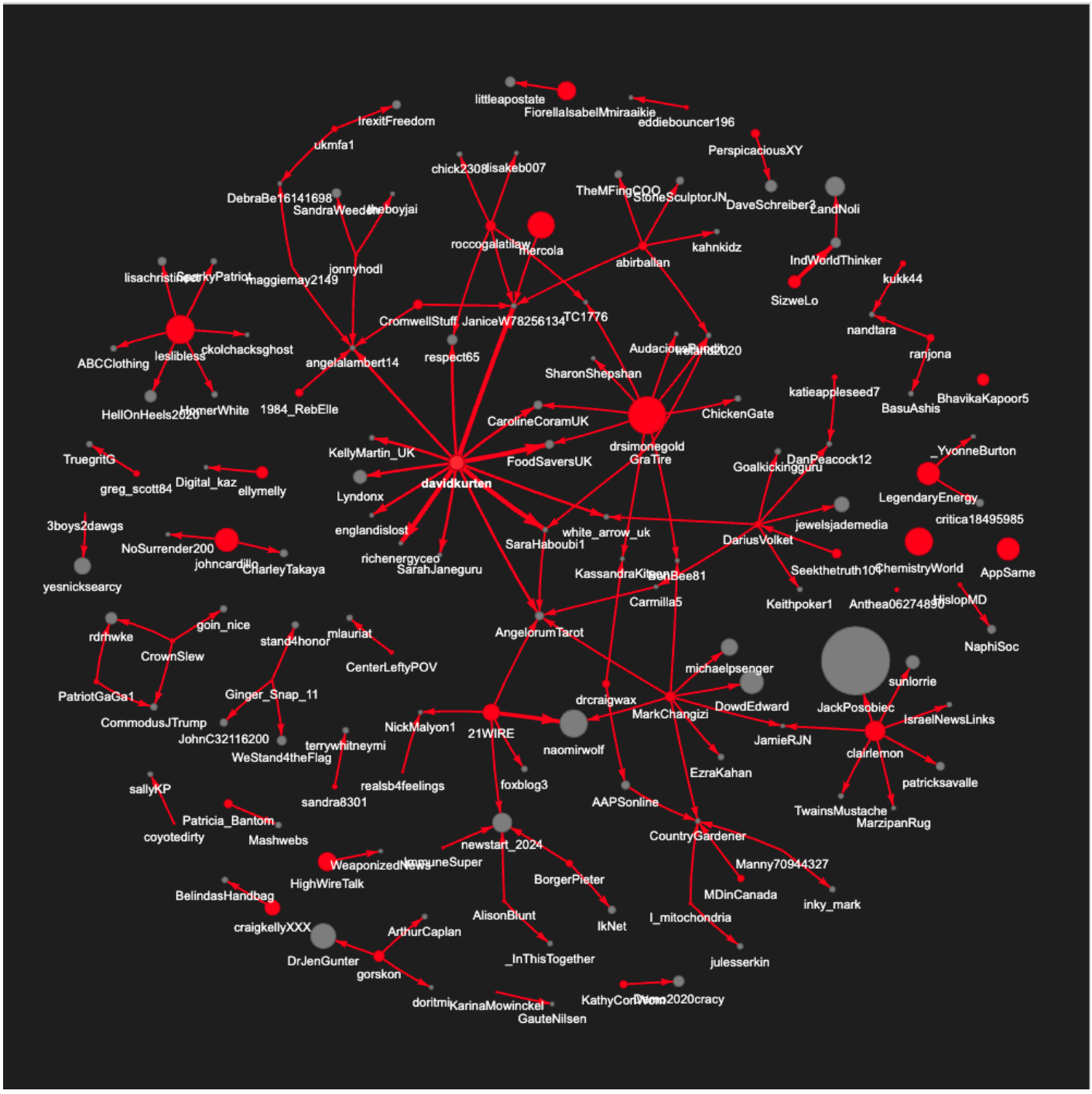


Figure 4. Network of the most impactful misinformers, along with retweeters having at least 10,000 followers. [Link to interactive graph](#).

DISCUSSION

When I started this study, I had failed to understand that publicly-available datasets of tweets contain only the tweet IDs, and that you have to pay X for the right to “rehydrate” the tweets, i.e. download the tweet text itself and any metadata. Because of this, I had to limit myself to a much smaller part of the set of 15M unlabeled vaccine-related tweets than I had initially intended to use. My budget allowed me to download 43,983 tweets from the three datasets I’ve worked with, along with information about 43,874 users. This particularly affects the network analysis, which would have benefited from being based on a larger set of tweets.

Another consequence of X’s prohibition against publicly sharing datasets containing full tweets is that it is impossible to fully reproduce earlier studies, whether to check their validity, or to keep building on them. Out of the 15,000 tweets labeled by Hayawi et al.[2], only 10,709 can still be accessed from X. The remaining 4,281 tweets have either been deleted, or made inaccessible to the general public. So even though the labeled dataset is only three years old and was created and shared as a resource to be used by other researchers, only 72% of it is still accessible.

A second issue with the data is that the set of 15M unlabeled vaccine-related tweets seems to have a much lower rate of vaccine-misinformation than the set of (originally) 15k labeled tweets. Out of the 10,709 tweets that are still available from the labeled set, 3,825, i.e 35.7%, are labeled as containing vaccine misinformation. A manual inspection of a random sample of 100 tweets from the unlabeled set found only 3 that I was confident should be labeled as misinformation, and another 7 of which I was not sure, suggesting that the percentage of vaccine misinformation tweets in the unlabeled dataset is likely to be below 10%. As a consequence, the 29,187 tweets purchased from the unlabeled vaccine-related dataset didn’t contain at all as many misinformation tweets as I had hoped to have access to.

Worth reporting is also that I was not able to reproduce the performances that Hayawi et al. [2] reported for their BERT-based misinformation classifier, or that Cotfas et al. [3] reported for their RoBERTa-based vaccine-hesitancy classifier. I used RoBERTa for both classifiers, as that gave better results. But even after hyper-parameter tuning, my misinformation classifier achieved a test-set F1-score of only 0.935, substantially lower than the 0.98 reported by Hayawi et al. [2]. For the vaccine-hesitancy classifier, I achieved a 75.4% accuracy for the three-class classification, far below the 95.57% reported by Cotfas et al. [3]. Most likely, part of the reason for this decreased performance is that only 72% of the tweets in the vaccine-misinformation dataset, and 71.5% of those in the vaccine-hesitancy dataset are still accessible through X, giving my recreations of them models less data to learn from.

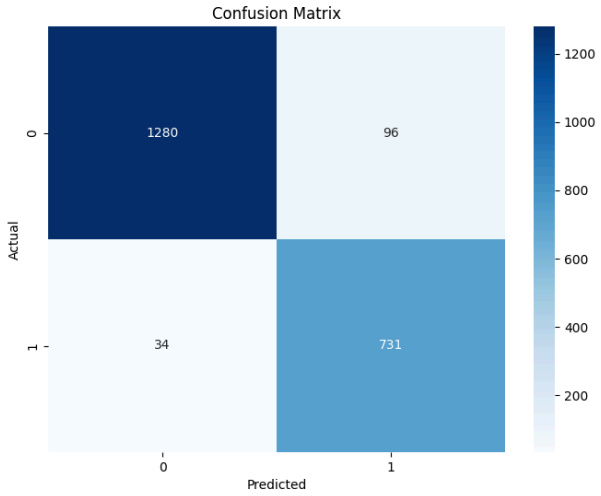


Figure 5. Confusion matrix for the vaccine-misinformation classifier. Accuracy: 93.9%. F1-score (Macro Average): 0.935.

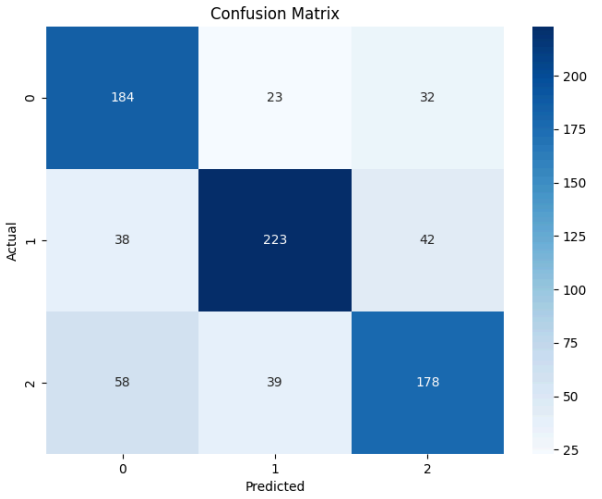


Figure 6. Confusion matrix for the vaccine-hesitancy classifier. (0: negative stance; 1: neutral; 2: positive). Accuracy: 75.4%. F1-Score (Macro Average): 0.714.

A related issue is that the vaccine-misinformation classifier seemed to fare much worse when applied to the unlabeled dataset than on the holdout test set from the labeled dataset. This is probably at least partly due to the much lower prevalence of misinformation tweets in the unlabeled dataset, as mentioned above. Out of the 29,187 unlabeled vaccine-related tweets downloaded, the classifier marked 51.7% as having a likelihood above 0.5 of containing misinformation, while my manual inspection, as mentioned above, suggested a prevalence of less than 10%. Raising the probability threshold for classifying a tweet as containing misinformation to 0.8, or even higher, lowered the rate of false positives, but the accuracy and the F1-score was too low for me to dare to use the classifier for the purpose for which I tried to recreate it: to allow me to extract a much larger set of misinformation tweets on which to conduct topic modeling and network analysis. The misinformation classifier fared somewhat better when applied to the tweets in the vaccine-hesitancy dataset, but still not well enough, based on manual inspection, for me to use the results. In the end, I limited myself to using just the labeled misinformation dataset.

As a consequence, I did not have enough tweets to conduct topic modeling on some specific subsets of the data that would have been interesting to analyze, e.g. tweets not containing misinformation, but still having a negative stance towards vaccines.

As for the network analysis, it would have benefited from being based on a much larger number of tweets. While the nodes in the graph are indeed all accounts that due to their large follower counts and/or prominent retweeters, have a large misinformation impact, the number of unique tweets displayed in the network graph is rather small. This makes the graph less robust against small changes than it ideally should have been. If we were to split the data in halves and construct two separate graphs, the nodes that currently are connected to many edges would still appear in both those graphs, but nodes connected with a single edge would be likely to appear in just one or the other of the graphs, indicating how sensitive this graph is to small changes.

CONCLUSION

In this project, I have extracted the main topics from a set of tweets containing misinformation about vaccines against COVID-19, and compared them against topics extracted from vaccine-related tweets that don't contain misinformation. Among the main misinformation topics were *Experimental gene therapy*, *Experimental and untested*, *Bioweapon*, *Depopulation*, and *Freedom and force*. Among the main non-misinformation topics were *Got the shot*, *Vaccinated and grateful*, and *Worry*. "Worry" was the most salient term in three out of five topics, even though they showed perfect separation in the Intertopic Distance Map, indicating just how important that term was in the non-misinformation conversation.

I have tentatively shown the relationships among the largest COVID-19 vaccine misinformers from December 2020 till July 2021, by organizing them into a network based on retweets. An analysis of this network revealed that just 15 accounts were responsible for just over 50% of the total vaccine misinformation impact represented in the dataset. These are the accounts that should be monitored most closely by anyone trying to counter the spread of vaccine misinformation.

Future work

A major limitation of the present study is the number of vaccine-misinformation that were included in the final topic modeling and network analysis. Work should go into developing a vaccine misinformation classifier with high accuracy on the kind of unfiltered data found "in the wild" on X. With such a classifier, it would be possible to extract larger sets of vaccine-misinformation tweets, which could then be used for topic modeling of subsets of interest, e.g. tweets not containing vaccine misinformation, but still showing a negative stance towards vaccines. A large set of vaccine-misinformation tweets would also make possible a more robust mapping of the most impactful misinformers and their relationships with each other.

REFERENCES

1. Mellis, C.; Lives saved by COVID-19 vaccines. Journal of Paediatrics and Child Health. 20 September 20122
2. Hayawi, K.; Shahriar, S.; Serhani, M.A.; Taleb, I.; Mathew, S.S. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. The Royal Society for Public Health. Published by Elsevier Ltd. 2021
3. Cotfas, L.-A.; Craciun, L.; Delcea, C.; Florescu, M.S.; Kovacs, E.-R.; Molanescu, A.G.; Orzan, M. Unveiling Vaccine Hesitancy on Twitter: Analyzing Trends and Reasons during the Emergence of COVID-19 Delta and Omicron

Variants. Vaccines 2023, 11, 1381.
https://doi.org/10.3390/vaccines11081381