Understanding Influential Accounts on Twitter

Aarush Shintre
aarush.shintre@gmail.com

ABSTRACT

Twitter is a popular microblogging application that has become increasingly popular for its ability to provide a communication platform for people worldwide, on various topics. Data from twitter is incredibly useful in understanding everyday people's opinions and sentiments in various contexts, including global issues, like the Covid-19 Pandemic, and the subsequent vaccination efforts. Looking at certain trends in sentiments, topics discussed and what popular influential accounts on twitter talk about are all great ways to understand what is relevant and important to people worldwide.

We tried to understand accounts' influence, by analyzing tweets spawned on twitter. First, we tried to understand when were covid-vaccine related tweets most popular in the time span of the data, and what factors might have affected popularity of these tweets. We categorized tweets into pro, anti, and neutral sentiments towards vaccinations, and further analysis was conducted on these categories.

Interestingly, we observed that about 42% of all favorites and retweets received throughout the time span were attributed to pro-vaccine tweets, which made up 30% of all tweets spawned. To understand what topics were discussed through these popular pro-vaccine tweets that helped them gain popularity, we conducted LDA Analysis on them. The pro-vaccine accounts provided information about covid vaccines, including discussions about vaccine efficacy and vaccine administration, which might have helped their tweets gain traction in the covid-19 vaccine discussion.

1. INTRODUCTION

In an increasingly technology-based world, various social media platforms provide a forum for discussion, and exchange of ideas. One such platform, Twitter, is a microblogging social media application. It is a powerful tool to understand a large population's beliefs regarding the vaccinations, and how they changed across time. Twitter has API's for easy access to large amounts of tweet data. Certain trends, as well as beliefs of the population can be understood from this data. We made use of twitter data to understand accounts' influence over Twitter by understanding their popularity, and the reach of their tweets, while also trying to understand what was discussed by influential accounts on twitter, in regards to the vaccinations/vaccination efforts.

The severe acute respiratory syndrome coronavirus 2(SARS-CoV-2) or Coronavirus was a worldwide pandemic, the first case of which was reported in December 2019. Efforts from various organizations led to vaccines being implemented from around December 2020. It was a widely discussed event, especially over social networking sites, like Twitter.

We looked at twitter data to understand an account's reactions and social interactions by performing numerical analysis. We utilized certain analysis methods to look at relatively large quantities of data to answer important questions such as: "How can influence be measured over twitter"? "What accounts were popular in this discussion"? "What do certain accounts who are popular over twitter talk about"? "How do sentiments towards covid-vaccinations change across time"? "How does the popularity of the tweets change across time"?

In recent years, various approaches have been made to answer questions from twitter

data, including some that overlap these. Methods have often included analyzing various trends over twitter by looking at not only the change across time of sentiments and opinions, but also popularity of both accounts and their tweets in various contexts. Additionally, previous work has looked into analyzing information spread, parts of the population that are involved in this process, and whether certain accounts, or groups of accounts tended to spread incorrect information or consistently negative information. Efforts have also been made to understand framing strategies used by popular twitter accounts, to understand how influentials gain popularity.

We focused to understand how sentiments towards vaccines changed across time, as well as the popularity of accounts making vaccine related tweets. Not only did this help understand overall trends of sentiment, but also what topics were being discussed. More importantly, it helped understand the popularity of accounts talking about different topics, with different viewpoints, and what was relevant/important to twitter users. This also helped us understand why certain accounts on twitter have a large audience and/or have great engagement through their tweets, and also understand the behavior and sentiment towards the vaccinations along with what opinions were most important to a majority of the accounts in context of covid vaccinations. We utilized the following approaches to do this:

1. Quantifying influence metrics, and measuring them to understand popularity for both influential tweets, and influential accounts.
2. Understanding the popularity of covid vaccine tweets across time, and also trying to understand how sentiments changed across time, and how they in turn affected popularity.
3. Understanding the topics discussed in tweets which were most popular.

2. RELATED WORK

Since the inception of social networking services, there has been a great opportunity to quantify influence and understand topic discourse and sentiments of the accounts in a relatively easy manner. With this, there has been a large amount of research in this field.

Previous research has worked to quantify influence and understand its power and emergence. Numerous methods have been implemented to understand influence, such as by implementing measures similar to PageRank (Weng et al. 2010), and looking at the factors behind retweet popularity (Suh, Hong, Pirolli & Chi, 2010).

Additionally, previous works have also looked into influence across a broad range of topics, so as to understand whether influence holds over topics, as well as what kind of influence is exhibited by these accounts (Blei, Ng & Jordan 2003); along with understanding the impact of these popular accounts, and checking whether information spread by them is truly accurate. (Lim, Toriumi & Yoshida 2021)

Further, previous research has looked into various aspects to understand the behavior of the influential accounts, some including understanding framing techniques used by them, along with modeling the topics discussed by them (Goldwasser, Pujari, 2021). All of these, among others, have enabled a greater understanding of interactions in the era of social networking.

Recently, research has also been conducted to understand Covid-19

vaccination discussions over twitter. Papers have looked into analyzing the sentiment of accounts towards the vaccines (Jang, Rempel, Roe, Adu, Carenini & Janjua 2022)(Bokaee & Mohammad 2022) , as well as sentiment based topic modeling (Huangfu, Mo, Zhang, Zeng & He 2022), among other techniques. Papers have also looked into vaccine hesitancy and spread of misinformation about covid-19 vaccines over twitter (Khan, Mallhi, Alotaibi, Alzarea, Alanazi, Tanveer & Hashmi 2020)(Thelwall, Kousha & Thelwall 2021), along with the impact of the misinformation(Loomba, Piatek & Larson 2021).

3. DATASET

We used Kaggle, a popular website for publishing datasets, among other data science and machine learning implementations, to access a Twitter dataset (Table 1).

Duration	Unique Accounts	Total Tweets
12 December 2020 to November 24 2021	85’550	228’207

Table 1: Dataset Description

The tweets in the dataset were about covid vaccines/vaccinations and started around the time that vaccines were initially announced, and span the time that their implementations on a wider scale were started. The biggest benefit of this dataset was that it spanned a relatively large time span, and included tweets related to only covid vaccines. The twitter accounts also spanned a large region.

The dataset includes numerous columns that provide information about both the account and the tweet, such as account name, the date the account was created, the location of the account, as well as the number of followers, favorites and friends of the accounts, along with text body of the tweet, hashtags included in the tweet body, and source of the tweet. It also included the date and time that a tweet was spawned, and how many favorites or retweets it generated.

4. Exploring Account Influence Metrics

To understand the influence over twitter, we looked at both, the influence of an account, as well as the influence of a tweet. We made use of certain metrics to conduct numerical analysis for both of these influence types.

4.1 Understanding how the account age affects the account impact

We decided to look at account age to see whether it was a factor in high influence; older accounts would have had established an audience and would have experience posting tweets that others would feel the need to share.

The dataset contains a user_created column, which indicates when an account was spawned. To understand account age, the date of each account being created was compared to the last date the tweet was made. To view distribution of popularity based on account age, an eCDF was drawn for three different categories: all accounts, top 10% oldest accounts, and top 1% oldest accounts, with respect to their cumulative probability of getting retweets (Fig. 1). An eCDF Plot is helpful to understand the distribution of the input variables, here, it is used to understand distribution of retweets

for three different categories discussed above.

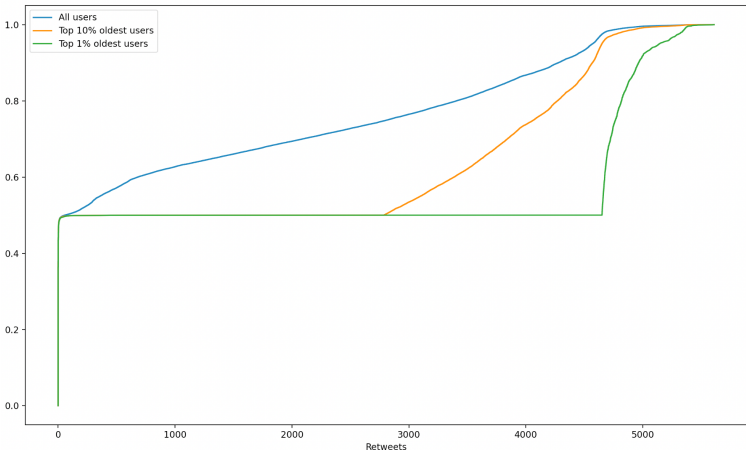


Fig. 1: Cumulative Probability of Retweet Popularity Based on account age

We observed that as the account age increases, the probability that the user receives more retweets also increases. About 80% of all the accounts received 4000 or fewer retweets. For the top 10% oldest accounts, about 80% received 4500 or fewer retweets and for the top 1% oldest accounts, about 80% received 5000 or fewer retweets. We assumed , with some certainty, that on average, older accounts tend to spawn tweets that receive more popularity.

4.2 Understanding the accounts influence

To understand the influence of twitter accounts, we looked to identify certain quantitative variables that would provide deep insight into a users' popularity. It is crucial to measure various factors, for example the account's direct audience, the amount of the account's idea propagation, and the engaging power of the account. Hence, we made use of three important metrics to measure **the influence of an account** :

1. *Indegree influence*, the number of followers of a user, directly indicates the size of the audience for that user. (Cha, Haddadi, Fabricio Benevenuto & Gummadi , 2010) (already present in the dataset as a separate column).

2. *Retweet influence*, which we measure through the number of retweets containing one's name, indicates the ability of that user to generate content with pass-along value. (Cha, Haddadi, Fabricio Benevenuto & Gummadi , 2010) (already present in the dataset as a separate column)

3. *Mention influence*, which we measure through the number of mentions containing one's name, indicates the ability of that user to engage others in a conversation. (Cha, Haddadi, Fabricio Benevenuto & Gummadi , 2010)

While the number of followers for a user and the total retweets a tweet got were already present in the dataset, total mentions a user receives is calculated from the tweet body. Mentions are present in tweet bodies in the form *@username.*

To understand how these three metrics vary, and whether some metric is not spread like the others, we plotted an eCDF Plot (Fig. 2). We noticed that almost all of the retweets and mentions are distributed lower than around 500; similarly the number of followers are distributed under 1000. This indicates that there are only a very small number of accounts who have a very high number of followers and tweets with very high popularity. A small number of accounts hold incredible influence as compared to the other accounts.

The traditional view assumes that a minority of members in a society possess

qualities that make them exceptionally persuasive in spreading ideas to others. These exceptional individuals drive trends on behalf of the majority of ordinary people; they are called the opinion leaders in the two-step flow theory (Katz and Lazarsfeld 1955), innovators in the diffusion of innovations theory (Rogers 1962), and hubs, connectors, or mavens in other work (Gladwell 2002) (Cha, Haddadi, Fabricio Benevenuto & Gummadi , 2010). This theory holds for this dataset.

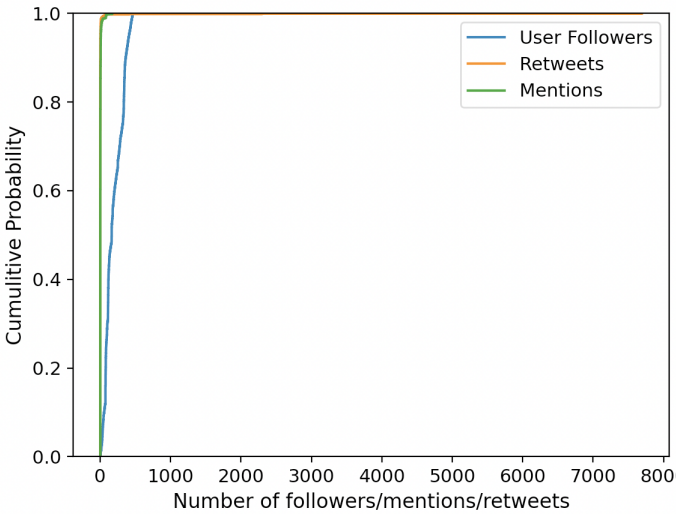


Fig. 2: Distribution of popularity metrics

4.3 Influential accounts with respect to Covid discussions

To understand what kind of accounts ranked highest in each of the three influence metrics, and whether certain accounts held multiple kinds of influence power; we looked at lists of top 10 influencers across each of the three metrics.

We found that the most followed accounts in the vaccination dataset were news sources (NDTV, CGTN, Times of India, China Xinhua News) as well as global organizations (WHO). These accounts, as suggested by the description, were directly connected to their audience. Hence, they had direct influence over their audience.

The most retweeted accounts, whose content people felt the need to share, included government officials (Dr. S. Jaishankar, Ananya Agarwal), influential people (Robert F. Kennedy Jr), and health organizations (U.S. FDA). The opinions of these individuals and accounts was relevant to other accounts, and hence, the information shared by them was propagated further, and in-fact, the most popular tweets could have even moved multiple tiers, influencing multiple accounts. A tier can be considered a level of accounts, so, for instance a certain tweet could be propagated from one user to another user, to another user, each of them forming a complex tier of accounts, each of them influencing others with the idea spawned by the root tweet. This would mean that a retweet could be immensely influential in spreading ideas and thoughts of a user.

The most mentioned accounts included global organizations (World Health Organization or WHO), national leaders, such as prime ministers or Presidents (narendramodi, potus), as well as vaccine manufacturers (pfizer, moderna). These accounts were extremely relevant and important to twitter accounts during this time period.

Interestingly, there was minimal overlap in the top 10 most followed accounts and most retweeted accounts as well as most mentioned accounts.

4.4 Understanding popularity of tweets

To further understand influence over twitter, we looked at the popularity of tweets along with that of the accounts:

a. Favorites a tweet received.

b. Number of times a tweet was retweeted.

Looking at these two popularity metrics over time helped understand at what point in time were covid vaccine related tweets gaining popularity (Fig 3). We observed that the first surge in covid-vaccine related tweet popularity was around March 2021, when a large number of nations (a few include India, Ghana, Ivory Coast) announced the first vaccine rollouts.

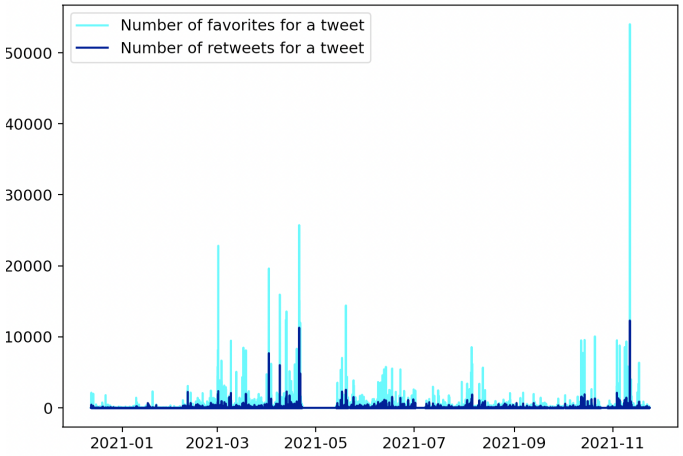


Fig. 3: Tweet Popularity over time

It might be interesting to understand what other factors might have caused this popularity, or how other factors might have affected popularity. To see this, we plotted a graph (Fig. 4) that shows:

1. Mean sentiment of tweets per month-both the mean polarity and mean subjectivity

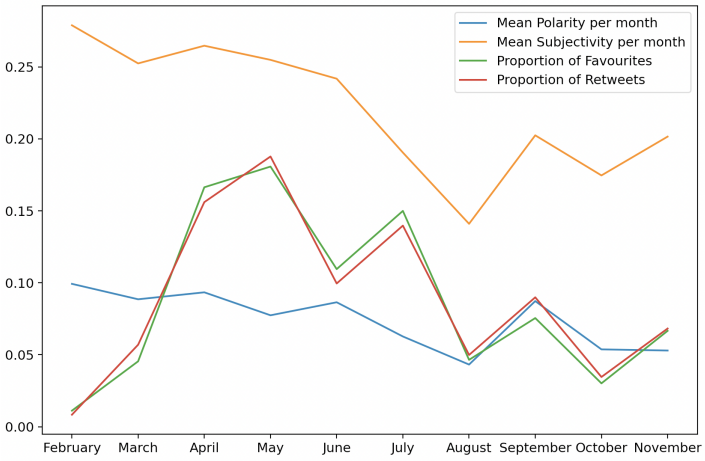
2. The proportion of retweets and favorites to all tweets made per month to the sum of retweets and sum of favorites made throughout the entire duration of the dataset, respectively. (represented by $p_{retweets}$ and $p_{favorites}$)

Normalizing by the total number of tweets posted on Twitter is essential to cancel out any variable effect on the data and allows the underlying characteristics of the data sets to be compared (Cha, Haddadi, Fabricio Benevenuto & Gummadi , 2010) .

$$p_{retweets} = \frac{Retweets\ per\ month}{Total\ retweets}$$

$$p_{favorites} = \frac{Favorites\ per\ month}{Total\ Favorites}$$

Both the polarity and subjectivity for each tweet were calculated using the TextBlob python module, which returns a polarity score ranging between [-1,1], where -1 defines a negative sentiment and +1 defines a positive sentiment, as well as a subjectivity score between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity meant that the text contained personal opinion rather than factual information.



g. 4: Proportional Popularity and mean sentiment p month

Figure 4 indicates that the mean polarity and subjectivity moved towards 0 until August 2021, after which it noted an increase. This indicated that accounts

spawning tweets until August were moving towards factual - instead of opinionated-tweets. Further, sentiment levels were dropping. Around July, as mean polarity kept decreasing, proportion of favorites/retweets per month dropped as well, meaning that covid vaccine tweets began receiving fewer popularity in respect to other time spans. This was seen until August, when as mean polarity and mean subjectivity scores began rising, proportional popularity of these tweets began increasing again. Higher positivity of tweets might have caused an overall rise in tweet popularity.

To understand this further, we calculated the proportion of anti vaccine, pro vaccine and neutral tweets per month. Pro-vaccine accounts were considered to be those whose polarity score was > 0, while the anti vaccine accounts were those whose polarity score was < 0, and neutral whose polarity score was 0. This data, along with tweet popularity (retweets and favorites) (Fig. 5) shows that for all months, a majority of the tweets are neutral, the next majority being pro vaccine accounts. Further, we observed that the proportion of pro vaccine accounts correlated with the overall popularity of tweets received to a great extent.

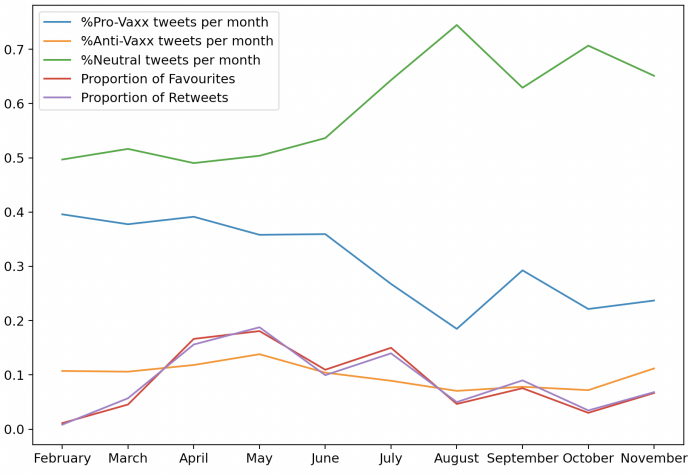


Fig. 5: Proportional Popularity and %account sentiments

Table 2 shows the percentage of retweets and favorites accounted for by these three types of accounts, along with the percent of the total accounts that actually held that sentiment. Interestingly, a smaller percent of pro vaccine accounts (about 29%) received nearly as many retweets and favorites as twice as many neutral accounts (about 60%) received. This indicated that positive vaccine-related tweets tended to be liked more. In general, pro vaccine accounts received more retweets and favorites. This also helped explain the correlation between the mean polarity score per month and the proportional popularity per month.

Tweet Sentiment	% Retweets	% Favorites	% Users
Neutral	48.51	48.1	60.54
Pro Vaccine	41.56	43.21	29.71
Anti Vaccine	9.9	8.7	9.74

Table 2: Distribution of Popularity Metrics

4.5 Understanding topics discussed by influential accounts

Since it was found that Pro-Vaccine tweets are generally more popular, we thought it would be interesting to see the words and topics these accounts implement to gain popularity.

4.5.1 Popular terms

First, we looked at the commonly used words in the tweets of both pro and anti-vaxx users. Looking at commonly used words can provide insight into the general idea that each of the two groups wishes to communicate.

We compared most commonly used words (excluding stop words like a, the, an, etc), along with their frequencies, in tweets made by Pro vaccine accounts, to that of Anti Vaccine Users (Fig. 6a & 6b). We noted that there was not much difference in the frequency of words for both types of accounts; both mention vaccines, vaccine manufacturers, etc.

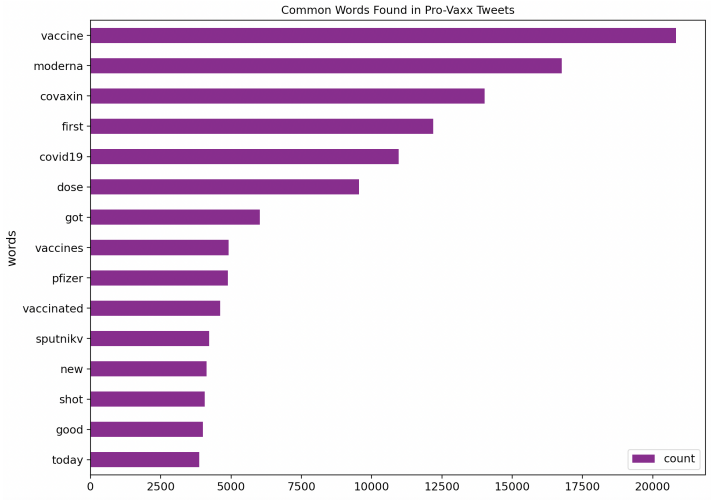


Fig. 6a: Common words occurring in positive twee

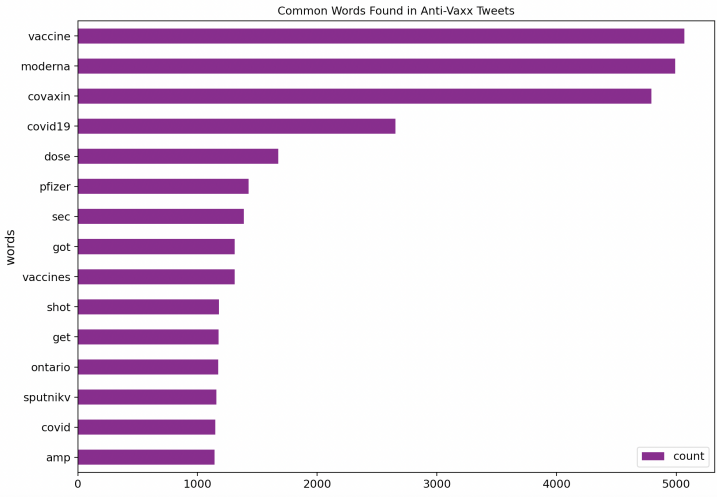


Fig. 6b: Common words occurring in negative tweets

4.5.2 Exploring discussed topics

Since the set of words used by both groups were quite similar, we felt it would be beneficial to see what context these words were used in, and what underlying topics were generally discussed through these Pro and Anti Vaxx tweets.

Pro-Vaxx accounts were immensely more popular than the Anti-Vaxx accounts, and understanding what topics they discussed might be interesting to understand what information they generally spread that other accounts found interesting and felt like propagating. It will also help reveal what topics were actually important to the majority of accounts. The underlying topics for both sets of accounts are shown in Tables 3 and 4, for anti and pro-vaxx accounts respectively.

LDA topic modeling was implemented to find what topics were generally used by accounts with anti-vaxx sentiments as well as those used by pro-vaxx sentiment accounts. Latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document (Blei, Ng & Jordan 2003). The Python gensim module, an open source library for unsupervised topic modeling uses modern statistical methods to generate an LDA model, based on pre-processed tweets.

Tweets pre-processing: First we pre-processed the tweets using lemmatization which involves converting a word to its base form. For instance, “apples” is converted to “apple”. Also, we lowered the

case of all characters, and removed links found in the tweets' body.

	Topic #1	Topic #2	Topic #3
1	expert	covaxin	moderna
2	subject	vaccine	booster
3	committee	usfda	shot
4	covid	sputnik	got
5	canada	amp	vaccine
6	cases	approve	dose
7	business	need	im
8	stop	home	get
9	russia	traditional	side
10	vaccine	wrong	pfizer
11	ontario	canada	vaccinated
12	evaluating	needs	covid
13	forced	health	second
14	kids	usa	nd
15	list	world	effects

Table 3: Topics used by Anti-Vaxx accounts

	Topic #1	Topic #2	Topic #3
1	moderna	covaxin	covid
2	dose	vaccine	vaccine
3	first	good	moderna
4	got	moderna	booster
5	vaccine	ages	pfizer
6	covid	ocgn	effective
7	shot	vaccinated	vaccines
8	pincode	free	sinopharm
9	slotsage	covid	man
10	today	people	new
11	get	pfizer	study
12	vaccinated	many	approved
13	covaxin	get	doses
14	im	news	safe
15	second	days	covishield

Table 4: Topics used by Pro-Vaxx accounts

The topics discussed by the two groups varied. While the Pro-Vaxx accounts made use of words like safe, slots, vaccinated, approved, first, second, got, etc, it indicates that these accounts discussed vaccinations, and their study results(like whether it was approved).

On the other hand, the topics discussed by anti-vaxx accounts included a wider array of words, like traditional, wrong, forced, business, stop, cases, effects, etc, that indicate the anti-vaxx accounts might have discussed other things like non-vaccine remedies, side effects, while also including words that indicate hesitancy, like forced, stop and wrong.

5. FUTURE WORK

It would be extremely helpful to have a dataset that tracks retweets, among other methods of information propagation over multiple tiers of accounts, in order to truly understand influence on opinions and sentiments. Understanding change in opinions as well as people reached by each of the retweeter will be crucial to understand the true magnitude of the root tweeters influence. Here, each tier represents a retweeter who is propagating a root tweet. A tweet could be retweeted by a certain first tier user, which, if retweeted again will reach more accounts. Further analysis can be conducted on these various tiers to

understand their opinions, as well as tweets spawned by them after this initial propagation, to understand how much they were influenced by the root tweet's opinions. Further, analysis can also be done on the retweeter's audience.

It would also be helpful to be able to find data for the popular accounts of the dataset, to see how their influence held over other topics, or was it limited to the subject of Covid-vaccinations.

6. CONCLUSION

We utilized data from a powerful social media application to understand how an account's opinions and behavior towards covid vaccinations were, and how their stance affected their influence and used various analytical methods to reach the conclusion that during the covid vaccination discourse over twitter, accounts with a pro-vaccination sentiment tend to have higher popularity than the other accounts. These accounts tend to discuss information about covid-vaccinations and covid-vaccine efficacy and testing that helped them gain traction during the duration of the dataset. The topics discussed by the far less popular anti-vaxx accounts were generally not related to covid-information and can be considered negative.

We found the most influential accounts in the covid-19 vaccine discussion for each of the three influence metrics. We also observed that older accounts tended to spawn tweets that were more popular. This could be explained by the fact that these accounts had an established audience, and maybe had experience posting content that was relevant to twitter users. We also identified the most influential accounts for each of the three metrics(indegree, retweet and mention), and observed the marginal overlap over the three influence metrics, which led us to the conclusion that type of influence power is independent of each other, and can be attributed to the type of account (Traditional News/ Government Body/ Politician/ Influencer/ etc.).

7. REFERENCES

1. Suh, Bongwon, et al. "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network." *2010 IEEE second international conference on social computing*. IEEE, 2010.

2. Weng, Jianshu, et al. "Twitterrank: finding topic-sensitive influential twitterers." *Proceedings of the third ACM international conference on Web search and data mining*. 2010.

3. Cha, Meeyoung, Hamed Haddadi, and Fabricio Benevenuto. "Gummad. KP Measuring user influence on Twitter: The million follower fallacy." *Proceedings of the fourth international aaai conference on weblogs and social media*. 2010.

4. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

5. Lim, Dongwoo, Fujio Toriumi, and Mitsuo Yoshida. "Do you trust experts on Twitter? Successful correction of COVID-19-related misinformation." *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2021.

6. Pujari, Rajkumar, and Dan Goldwasser. "Understanding politics via contextualized discourse processing." *arXiv preprint arXiv:2012.15784* (2020)

7. Jang, Hyeju, et al. "Tracking Public Attitudes Toward COVID-19 Vaccination on Tweets in Canada: Using Aspect-Based Sentiment Analysis." *Journal of Medical Internet Research* 24.3 (2022): e35016.

8. Huangfu, Luwen, et al. "COVID-19 Vaccine Tweets After Vaccine Rollout: Sentiment–Based Topic Modeling." *Journal of medical Internet research* 24.2 (2022): e31726.

9. Nezhad, Zahra Bokaee, and Mohammad Ali Deihimi. "Twitter sentiment analysis from Iran about COVID 19 vaccine." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 16.1 (2022): 102367.

10. Rogers, E. M. “Diffusion of Innovations” Free Press, 1962

11. Katz E, and Lazarsfeld P “Personal Influence: The Part Played by People in the Flow of Mass Communications” New York: The Free Press. 1955

12. Gladwell, M “The Tipping Point: How Little Things Can Make a Big Difference” Back Bay Books, 2002

13. Khan, Yusra Habib, et al. "Threat of COVID-19 vaccine hesitancy in Pakistan: the need for measures to neutralize misleading narratives." *The American journal of tropical medicine and hygiene* 103.2 (2020): 603.

14. Loomba, Sahil, et al. "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA." *Nature human behaviour* 5.3 (2021): 337-348.

15. Thelwall, Mike, Kayvan Kousha, and Saheeda Thelwall. "Covid-19 vaccine hesitancy on English-language Twitter." *Profesional de la información (EPI)* 30.2 (2021).