Analyzing Causes of Vaccine Hesitancy through Twitter Communication
Comp 598 Project

Robert Saad[1], Massimo Vicenzo[2], Zachary Vernec[3]
McGill University[1,2,3]
robert.saad2@mail.mcgill.ca[1], massimo.vicenzo@mail.mcgill.ca[2], zachary.vernec@mail.mcgill.ca[3]

Introduction

Our team has a goal to understand the discussion around COVID-19 happening recently. This pandemic has been playing a huge role in the world's functioning, as well as generating a societal conversation on how to mitigate its consequences. In most of the English-speaking world, we are now at a point where most of the eligible population has been vaccinated, and the conversation has shifted to how to reduce vaccine hesitancy. Accordingly, we have developed a typology for vaccine-related English-language tweets to determine which topics generate the most anti-vaccine and anti-vaccine measures sentiment and what those topics consist of.

Our team collected and annotated 1,000 original tweets using the Twitter API and filters to narrow results to our research interest. We conducted an open coding on the first 200 tweets and obtained a two-axis typology to categorize Covid-19 vaccine-related tweets, obtaining: Scientific, Sociopolitical and Economic vs Opinion, Reaction, Information. In addition, using engagement metrics such as the number of likes, retweets, and replies, we were able to locate the kind of content that played an important role in increasing or decreasing the fear around the vaccine, and using the TF-IDF method we were able to better characterize what was being discussed in each topic in more detail.

Our analysis led us to three main aspects. First, as surprising as it is, the economic causes and consequences that led to vaccination are not often mentioned throughout our data set. Then, the scientific aspect of the vaccination campaign is an important source of the hesitancy, however, the apparition of the omicron variant may have changed the results of what information is posted on a day-to-day basis on Twitter. So, our interpretation was particularly focused on the consequences of the political opinions and reactions posted online. The increased negativity read on social media, is what today has led to the increase in vaccination hesitancy. Our team is persuaded that this analysis reveals just the tip of the iceberg, and the question of vaccination hesitancy is more complex.

Data and Data Collection Process

To find tweets relating to our topic, we queried the Twitter Core API, using a list of case- and punctuation-insensitive keywords the tweets must contain. These keywords can be broadly categorized into two groups: vaccine-related words and vaccine brand names. The former category consisted of words where vaccine is the root of the word (for example "vaccinated"), but also more informal expressions such as the "vax"-derived words, the "jab", and the needle emoji. These informal keywords are important in making the sample representative, as formality is heavily correlated with social context, and therefore likely with the kinds of messages being shared. For the vaccine brand keyword category, we included the proper nouns commonly being used to refer to specific COVID-19 vaccines, such as Pfizer, BioNTech, and Spikevax. We also made sure to include abbreviations and alternate spellings for certain brands, for example "Johnson and Johnson" could also be written as "Johnson & Johnson" or "J&J".

In general, our keywords needed no special attention, as the nature of a pandemic is that the main vaccines / pharmaceutical products from companies that are currently being talked about are those relating to COVID-19. However, there were some occasions where we found we needed to exclude keywords from our query. For example, when searching for Novavax-related tweets, we had to exclude those with its stock symbol NVAX as share value was changing for a reason unrelated to the company's COVID-19 vaccine.

On the other hand, we have chosen not to include "COVID" and "pandemic" as keywords to search for. While we had considered them a possibility, we soon realised this caused us to get tweets which, while they would get tweets and opinions relevant to COVID, were not necessarily about vaccines, and so we deemed them an ill fit to answer our question of vaccine hesitancy.

Furthermore, we disallowed “quote tweets”, “retweets” and “replies” to be in our pool of tweets to pick from. Retweets and replies posed a problem, since they would cause us to potentially pick the same tweets more than once, and we wanted people to see people's own original opinions, not those they formed from others on the platform. The problem with quote tweets was different: because the Twitter API would find the keywords within the tweet that has been quoted, as well as the tweet itself. This was causing tremendous slowdowns during the annotation phase, since if we had a tweet that did not have any keyword, we would have to go to the actual tweet on the platform to see what the context of the quote tweet was.

And so, with these restrictions, we gathered 1000 tweets, with an additional 500 extra to replace any false positives matches (unrelated to vaccination), which amounted to just being 49 tweets. Along with the text of the tweet, we also gathered the unique ID of the tweet (for future reference), and the engagement metrics: like, retweet, and replies count.

Methods

To develop a typology, all team members conducted the first step of an open coding individually, before discussing which of those three tweet categorization schemes would be most useful to continue with. After briefly defining our chosen typology, we then continued our open coding individually, stopping after each pass through our subsections of the dataset to clarify edge cases in our definitions. After 3 passes, we had satisfyingly clear-cut definitions, which we used to annotate the remainder of the data. The final typology we used will be explained in the Topics subsection of our results.

In parallel to developing a typology of topics, we had to define a proper measure of sentiment to help understand the data in relation to vaccine hesitancy. While a common measure in sentiment analysis is based on the phrasing of the tweets in terms of positively/negatively correlated use of words, we chose to specifically focus on user’s sentiment towards vaccination and vaccine-related measures, no matter the terms being used to express that sentiment. For example, a user being annoyed at others not getting vaccinated would still be considered to have positive sentiment, while a user praising a lawsuit against vaccine mandates would still be considered to have negative sentiment. Note that to remain unbiased and keep our sentiment analysis clear and consistent across annotators, all vaccine measures were considered under the same umbrella, no matter how extreme. In addition to positive and negative sentiment, we included a type for neutral sentiment, as many tweets relating to vaccination took no sides on the issue, and others tried to consider both sides at the same time. This clear and simple ternary typology allowed us to complete the annotation process quickly, avoiding mistakes which can be caused by fatigue.

For our main analysis of sentiment and tweet reception, we focused a large part on the most basic of statistical tools such as counts, averages, and averages of non-zero values. For sentiment analysis, we coded positive/negative/neutral sentiment with +1/-1/0 respectively, which allowed us to additionally combine sentiment of different types to get an overall picture. As opinions are not really something that could be perfectly described by 3 numbers, it would have been more accurate to have this value be expressed as any real number between 1 and –1, which is what we originally anticipated doing. However, this type of annotation was unreasonable. Quantifying how much someone supports a certain topic by hand is a difficult thing to do, especially since we could ask, “what is the difference between a rating of 0.50 and 0.51?”, so we decided that a discrete rating system would fit better for our methods. Even with a discrete rating system, since our entire team was participating in the annotation process, personal bias could be a factor in what rating a tweet receives. To counteract this, we reduced our rating system to the ternary system described above; regardless of the degree of how in-favour or not in-favour a tweet may seem, it can only be given a 1 or –1 respectively, or a 0 if it shares no opinion or both opinions.

In contrast, to get an accurate description of the categories we established with our typology, we used some more complex machinery. While the basis of our descriptions is to use TF-IDF to extract representative words from each category, we used the industry-ready package spaCy to process our tweet data in the appropriate format. Using its basic pretrained English model, we were able to tokenize in a more context-dependent manner than naïve rules-based methods, which was useful to prevent splitting of entities in some cases (e.g., “U.K.” being kept as a token representing the United Kingdom). Our other use for the spaCy pipeline was to lemmatize most words to simpler lower-case versions, since for our TF-IDF we are more interested in getting a general feel for the topics rather than a specific word used. In our cleaning pipeline, we made sure to filter on stopwords and punctuation-only tokens, as is standard. In addition, we were forced to filter out emojis, since our results were getting biased by single tweets which used many repeating emojis. A last point of note in this data processing is that spaCy’s smaller model had difficulties making the difference between words in title-case and proper nouns, and so kept some words capitalized. With our time constraints, we were unable to work around this issue using a larger model, nor were we able to convert words to lowercase as the benefits would be outweighed by the information lost about entities, in our opinion.

Results: Topics

After conducting our open coding, we decided that we would have two axes across which to categorize our tweets.

First, we would identify what was the primary subject of the tweet, that is, what the tweet was about: scientific/health-related, sociopolitical, or economic. Note that while "economic" was underrepresented in the tweets used for our open coding and might not have been considered as its own topic under different circumstances, we thought it important to include as pre-vaccine this was a noticeable driver of the anti-measure conversation (regarding lockdowns).

The other axis we looked at was the information source of the tweet, that is, where the underlying information is coming from: news of an event/informational, reaction to specific event, or a standalone statement/opinion. Combining these two axes gave us nine different topics we used to categorize the tweets while allowing us to disentangle hesitancy-related information, as we hypothesized that there is a meaningful difference between where the vaccine hesitancy is coming from and towards what subject it is being articulated.

The following are the full definitions obtained from our open coding, with examples for the actual topics afterwards.

Definitions

Scientific/Health-related: Any tweet which focuses on medicine (including vaccine efficacy, vaccine symptoms, covid symptoms), epidemiology (including contact-tracing and variants) and demography (including vaccination rate). Scientific rigour and veracity of the claims are not considered.

Sociopolitical: Any tweet which contains discussion about the government, the way it is run. Laws including vaccine mandates go here when talked about in the abstract. The social aspect of this category also includes any general views on what should be done that do not include scientific reasoning, discussion about celebrities (including death from COVID-19), or discussions about the media. Claims about personal vaccination status are also included, as they are implicitly a social pressure to others getting vaccinated.

Economic: Any tweet which is related to specific loss of jobs, trade, commerce, prices, stocks.

News/informational: Any tweet from an organization that aims to inform, without an individual's commentary. Also, any tweet from a user whose sole act is pushing a piece of information or article without include additional commentary alongside it. The validity of the news is not taken into consideration, nor is its level of bias.

Reaction: Any tweet either discussing a specific recent event or sharing a source of news alongside the user's own opinion.

Opinion/Statement: Any tweet stating an opinion without reference to a specific event. Included in this category are tweets describing someone's action/non-action, such as someone stating: "I went out and got my booster shot today."

Examples

Scientific News:
Tweet ID: 1466130310710579207

> "Vaccinated California man first in US to test positive for Omicron – Times of India https://t.co/PuW31Q6kiS"

Scientific Reaction:
Tweet ID: 1466129751001772035

> "Did I read that correctly? Omicron has been detected in California by someone that was fully vaccinated? So why is something that isn't very effective still being mandated?"

Scientific Opinion:
Tweet ID: 1466129930077585408

> "My cousin had his Booster shot today and then tested positive for covid. Now you wonder why they didn't wait for test results before giving him nxa"

Political News:
Tweet ID: 1466129900654501900

> "A group of Republicans could briefly shut down the government over vaccine mandates - BuzzFeed News https://t.co/ut1QZrNZqK"

Political Reaction:
Tweet ID: 1466130350346850309

> "Jen Psaki is lying again today during her press conference saying Trump told people to inject bleach. That's BS. She also said unvaccinated people get the China virus more easily than vaccinated folks. Tell that to all the folks who had both shots and still got the Vid."

Political Opinion:
Tweet ID: 1466130590755872783

> "Control the world thru religion and vaccination."

Economic News:
Tweet ID: 1466128528534306816

> "Hundreds of NYC jailers face suspension over vaccine mandate - It was delayed a month for jail workers because of existing staffing shortages.Jail workers

> who’ve applied for religious or medical e... - https://t.co/6BMNEp489J https://t.co/SBJov3dwGW”

Economic Reaction:
Tweet ID: 1466129946917679104

> “This is so wrong at so many levels â€¦ PAYING people unemployment $ who quit their job because they won’t get #COVID19 vaccinated.
>
> This is our DAM taxpayer $ that @GovRonDeSantis is wasting! He is beyond contempt. This is a tax hike for all Floridians! https://t.co/La8JDgYig5”

Economic Opinion:
Tweet ID: 1466129604855353352

> “Our jobs shouldn’t be on the line for this damn vaccine either.”

Findings

	Scientific	Political	Economic	All
News	14.0	20.3	2.2	36.5
Reaction	10.2	13.7	0.4	24.3
Opinion	15.1	22.7	1.4	39.2
All	39.3	56.7	4.0	100

Table 1: Distribution of Tweets in Chosen Topics

Results: Engagement and Reception

To get the further insight into which topics relate more to vaccine-related hesitancy, we wanted to measure engagement and the reception of the tweets that we had gathered. We needed this to see what topics get the most attention, as original tweets are only a single component of how users interact with twitter, and reception?

Engagement

To analyze engagement, we used the amount of “likes”, “replies” and “retweets” each tweet had as metrics. Excluding quote tweeting, these three metrics are the only ways people can engage with a tweet. We did not include quote tweets as the three other ways of interaction on twitter are much simpler for the average user to use and so we expect there to be less non-zero values for this metric, and while a quote tweet is interacting with the original tweet, some might consider it less as an interaction and more of a reaction, while a reply is directly interacting with the user who posted the original tweet.

Some things to keep in mind when looking at these numbers were that first, many tweets do not get that much interaction. For most users, their posts are rarely viewed by anyone, let alone liked or replied to. This phenomenon of a heavily skewed distribution was present in our data, as Table 2 clearly demonstrates.

	Retweets	Replies	Likes	All
= 0	85.8	87.1	75.1	69.5
< 5	97.4	98.3	93.3	93.0
< 10	99.0	99.4	96.5	96.5

Table 2: Percentage of Tweets with a Small Amount of Interactions

Also, something to note is that we cannot treat all metrics equally: while likes and retweets can clearly determine that people support this tweet, things are not so clear when it comes to replies. In fact, if there is a high reply count compared to the like/retweet count, this usually implies that a tweet was controversial in some sense. Of course, since we are looking at tweets from absolutely anybody, this ratio might not be as accurate since most tweets don’t receive much interaction.

Therefore, to better understand the engagement those tweets had and how they impacted vaccine hesitancy, we will have to take these things into account when performing our analysis.

Reception

As stated in the Methods section, we used a ternary system where negative, neutral, and positive reception were given the value -1, 0, or 1 respectively. Specifically, being of a certain reception, means that vaccines or vaccine measures are being discussed in a positive, neutral, or negative way.

Findings

Tables 3 through 5 show our findings for the engagement with our topics.

First, looking at Table 1, we can see that the economic topic was a very small portion of our data, which implies that the values for engagement for the economic topics are possibly skewed due to the small sample size. For this reason, we will avoid making any assumptions about the engagement on these topics, and resort to discussing the possible reasons for this phenomenon in our later analysis. If we do exclude economic topics, we see that order of topics by amount of engagement is very roughly preserved as we compare between metrics including or excluding tweets with 0 of that interaction, with topics usually only moving 1 position up or down in each ordering. It is for this reason that we will not distinguish between both cases in our analysis.

	Scientific	Political	Economic	All			Scientific	Political	Economic	All
News	2.24	1.05	2.09	1.57		News	11.63	4.98	15.33	7.86
Reaction	1.43	1.45	1.50	1.44		Reaction	4.87	7.07	6.00	5.93
Opinion	2.83	1.54	0.29	1.99		Opinion	9.95	4.92	1.33	6.68
All	2.26	1.34	1.40			All	8.88	5.36	8.0	

Table 3: Average Likes in Chosen Topics
Left: Average Likes including Tweets with 0 Likes | Right: Average Likes excluding Tweets with 0 Likes

	Scientific	Political	Economic	All			Scientific	Political	Economic	All
News	0.76	0.50	0.41	0.60		News	5.10	2.91	4.5	3.76
Reaction	0.59	0.49	0.25	0.53		Reaction	2.61	3.72	1.0	3.45
Opinion	0.59	0.24	0.07	0.37		Opinion	5.24	2.29	1.0	3.05
All	0.65	0.40	0.28			All	4.20	2.91	2.75	

Table 4: Average Retweets in Chosen Topics
Left: Average Retweets including Tweets with 0 Retweets | Right: Average Retweets excluding Tweets with 0 Retweets

	Scientific	Political	Economic	All			Scientific	Political	Economic	All
News	0.56	0.28	0.14	0.38		News	4.49	2.55	3.00	3.54
Reaction	0.21	0.37	0.25	0.30		Reaction	1.40	2.68	1.00	2.09
Opinion	0.37	0.27	0.07	0.30		Opinion	2.55	1.91	1.00	2.15
All	0.40	0.30	0.12			All	2.94	2.3	1.67	

Table 5: Average Replies in Chosen Topics
Left: Average Replies including Tweets with 0 Replies | Right: Average Likes excluding Tweets with 0 Replies

	Scientific	Political	Economic	All
News	0.24	0.21	0.14	0.22
Reaction	0.30	0.20	0.25	0.25
Opinion	0.43	0.39	0.21	0.40
All	0.33	0.28	0.17	

Table 6: Fraction of Positive Reception in Chosen Topics

	Scientific	Political	Economic	All
News	0.50	0.62	0.50	0.57
Reaction	0.15	0.22	0.25	0.19
Opinion	0.07	0.20	0.21	0.15
All	0.24	0.36	0.38	

Table 7: Fraction of Neutral Reception in Chosen Topics

	Scientific	Political	Economic	All
News	0.26	0.17	0.36	0.22
Reaction	0.55	0.58	0.50	0.56
Opinion	0.50	0.41	0.57	0.45
All	0.43	0.36	0.45	

Table 8: Fraction of Negative Reception in Chosen Topics

Table 6 through 8 show the results we gathered from our reception annotation. Similarly, to what was discussed above, since the economic category was small compared to the size of our sample, the values that we have for the economic topic should not be considered a perfect representation of the domain studied, due to this small sample size.

Since our reception annotation takes 1, 0 and –1 as values, we can calculate what we call the reception score, which is just the average value that each of the topics has for reception. This can be seen in Table 9 which follows.

	Scientific	Political	Economic	All
News	-0.03	0.04	-0.23	0.00
Reaction	-0.25	-0.37	-0.25	-0.32
Opinion	-0.07	-0.01	-0.36	-0.05
All	-0.10	-0.08	-0.28	

Table 9: Reception Score in Chosen Topics

Results: TF-IDF

Term Frequency - Inverse Document Frequency, or TF-IDF, is a method we used to further analyse each of our topics. Each word is given a score based on how often it appears in a topic, this score is then reduced by how many topics this

Topic	Top 10 Words with the Highest TF-IDF Scores
Scientific News	South, California, variant, Nov., symptom, Africa, feature, booster, person, U.S.
Scientific Reaction	variant, person, Fauci, South, quarantine, California, booster, OmicronVarient, conference, wtf
Scientific Opinion	booster, collapse, fan, vaccine, UNVACCINATED, WATCHE, ill, mutate, effect, feel
Political News	Mandate, Republicans, Biden, shutdown, Lamb, Dies, Network, government, Marcus, Government
Political Reaction	Biden, Jen, Psaki, court, THREATENING, mongering, prison, sense, vax, @news24
Political Opinion	booster, entire, society, hate, Covid, choose, happy, useless, vaccine, abortion
Economic News	Guard, Lloyd, MRNA, cop, Austin, Defense, National, loss, lose, market
Economic Reaction	Business, Centene, DAM, Floridians, G20, Job, Johnson, Morgan, Stanley, UPS
Economic Opinion	asset, sell, supply, profit, @LumenTechCo, @WolframResearch, @daktari1, @usps, ABUS, BCpoli

Table 10: TF-IDF Results
TF-IDF of a word was given by: {# of times the word occurs}*log({# of topics}/{# of documents the word appears in})

word appears in. In essence, the TF-IDF method should allow us to extract somewhat unique words for each topic. Given these words, we can try and extrapolate the main discussion of each topic. The exact formula we used, and our top ten results are in the following table.

Discussion

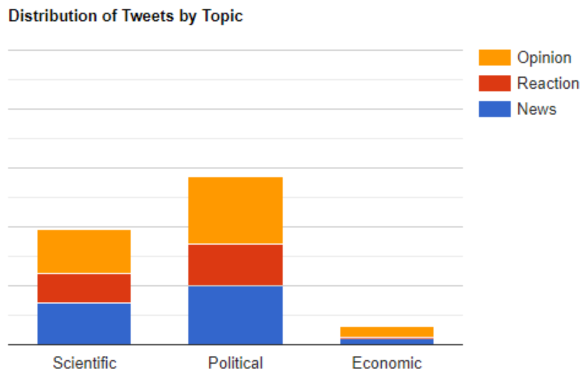
We conduct our analysis separately on each of the different results we found, and later we conclude with our thoughts on what our results show overall, with respect to the original question.

Distribution of Tweets

Looking at the distribution of tweets in our sample, we can extrapolate two main points. First, we can see that economic discussion is not a topic that surrounds vaccines, and so cannot be a main concern of vaccine hesitancy. We found this to be unexpected, as economic considerations were an oft repeated argument against early lockdowns (a pre-vaccine measure) and so we had expected a certain number of tweets discussing how the vaccine mandate was causing people to lose their jobs or force them to get vaccinated. It is possible that vaccine mandates aren't being fully enforced everywhere and so this would not be a cause for concern for many people. Nonetheless, since this portion of the dataset is small, doing analysis on engagement and reception might not be wholly accurate.

In fact, while economic discussion was low, political discussion was much higher than we anticipated as well. While we did anticipate a higher percentage to be occupied by political tweets then there should be for a less politicized health crisis, we did not expect it to have more than the quantity of tweets discussing the vaccines, symptoms, and actual sickness. We concluded that this is simply from people turning a vaccination into a question of rights and mandates, but we will investigate this further in the TF-IDF section.

As for the other axis of categorization, we found no peculiar results. Opinion and News tweets were among the most common, which isn't surprising as Twitter is a platform for a user to share their own thoughts and actions, and as the COVID-19 pandemic is a global crisis, we would expect a lot of articles and news reports to be written, and subsequently promoted.



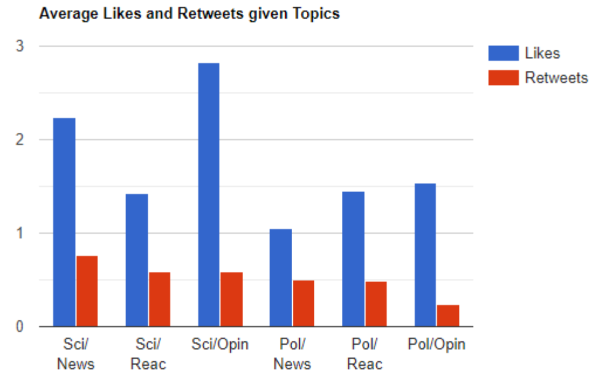
Graph 1: Distribution of Tweets by Topic

Engagement

We have shown before that most tweets do not have any engagement whatsoever (Table 2), and so we will focus on discussing on where the most engagement happens rather than the actual quantity.

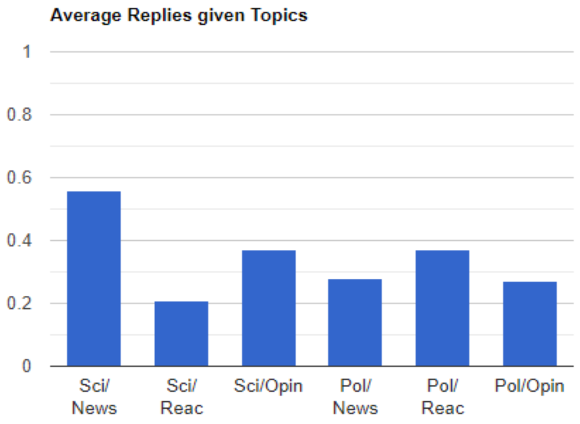
Comparing the values for average likes and retweets we see that most of these types of interactions are in scientific news and opinions topics. The interaction with news sources is understandable, as these actions can spread a tweet to a larger audience, which would be the goal of most people when liking or retweeting a news article. Meanwhile, opinions are highly liked, presumably since users would like other user's opinions if it resonated with them or they agree.

In graph 2, we can also notice that the fact that there are more tweets, but fewer likes in the politics category could be attributed to the simple fact that politics are not a subject that one can agree with completely. While getting a vaccine or not is a binary choice, politics are rarely as clear cut.



Graph 2: Average Likes and Retweets by Topic

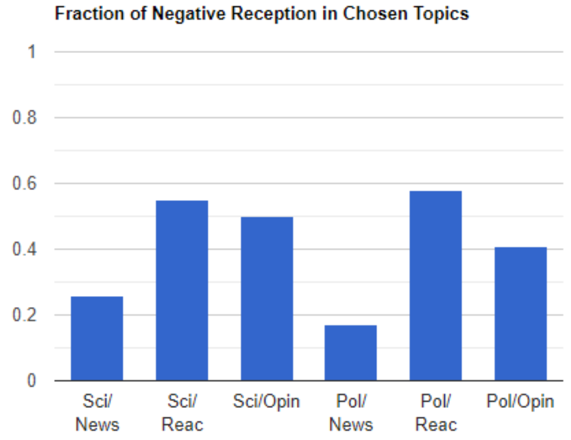
Replies, as we expected, do not follow the exact same trend. Of course, if a tweet has a lot of one type of reaction, we expect many of all types of interaction, however compared to likes and retweets we can still see a difference. Politics averages the same number of replies as the scientific topics. This makes theoretical sense, as politics is a topic that sparks more debates, but seeing as scientific topics have more interactions in general, would mean that it should also have a high reply count as well. This is what we are assuming causes the overall similarity between the average replies.



Graph 3: Average Replies by Topic

Reception and TF-IDF

An important way to analyze the data would be to focus on one main aspect: the negativity expressed per topic compared to the number of tweets posted on those topics. We assumed that if we were to find the root cause of vaccine hesitancy, it would appear in tweets that had a negative reception towards vaccines. So, our last decision was to look at the average scores or the fraction of negative responses per topic. We decided to put our attention towards the topics with the highest number of negative responses. The reason for this was that the reception score being close to 0 does not necessarily mean there is not a lot of negativity/vaccine-hesitancy, as the topic could also just be very divisive and have also many positive responses. Therefore, we will consider topics with high fractions of negative responses as sources of hesitancy.



Graph 4: Fraction of Negative Reception in Chosen Topics

As explained previously, due to the small number of economic-related tweets, we will discard it from the analysis. Therefore, when looking at Graph 4, we notice that reaction and opinions are leading the scores in negativity meaning that whatever the topic is, the hesitancy is transmitted from peer to peer. Reaction and opinion are close in the sense that both convey non-objective information. Regarding the reactions, the political topic is leading and in terms of opinions.

Now it is interesting to look at the results of the TF-IDF section of our code to analyze what could be the main source for the political and scientific negativity. First, for the scientific reaction, we notice an unusual word: WATCHE. Inspecting our data, we find that it is a hashtag (alternatively stylized #WatChe) standing for Watford vs Chelsea, a soccer game where players and fans kept fainting and collapsing which some have attributed to the covid vaccine. This also informs our interpretation of the other words in that category, such as “fan” and “collapse”. Another important piece of understanding revealed by the TF-IDF results in this section is that Omicron variant is one the most crucial factor. With a lot of people questioning the vaccine since the person who brought over the variant to the United States, was in fact vaccinated.

These both can be considered outliers in our dataset. Since these are sizable events, this could have possibly caused a shift in the distribution of tweets, and of course the reception of the vaccine. However, we would hypothesise that from the medical side that the issue with vaccines mainly comes from vaccine efficacy. The Omicron variant has seemed to strengthen the belief that the vaccine is useless, using the first variant-infected person as an example. The validity of this claim is not considered. Of course, this is what people are discussing now, when the new variant has just appeared, if the sample was taken just days before, we may have possibly found different, more general point of view. Although we do think this result is interesting and pertinent.

Regarding the political opinions, we notice an interesting trend with the use of words like “hate”, “choose”, “useless”, “society”, “abortion”. Users whose tweets fall in this category, seem to challenge the efficacy of the vaccine, or seem to be using this to further beliefs about rights. Specifically, they seem to be feeling attacked, as they feel they are being forced to take the vaccine, which would be an infringement on their right to choose how they live. Funnily enough, not only would this discussion be used against vaccination, but it is also used as an argument for abortion. The word “abortion” was surprisingly common, and it was because people would often compare the lack of choice with getting a vaccine, with the current lack of choice some women have with getting an abortion. Of course, this is unrelated to vaccine hesitancy, however it shows that at the very least, users are using the pandemic and more specifically vaccination to discuss rights and privileges.

The scientific results we got from the analyses can be skewing our results. Indeed, the discovery of the new variant at the time of the data collection may have shifted our scientific results from daily reality and therefore may be used with precautions. On the contrary, the results of the analysis related to the political section give us the results we were expecting. It seems that the lack of choice and freedom in the mandates is what is creating a huge negative wave towards the vaccine, more than the medical point of view. We can conclude that the main reason for the hesitancy seems to be political. The news plays a significant role of conveying information that is neutral from our annotation, however, the way the population interprets them, reacts to them, and how they convey their own opinions is really affecting the hesitancy to get vaccinated. We cannot assess what is the specific political source, but we can interpret that the way the vaccine is politically handled and how the mandates are affecting people’s freedom makes the hesitancy increase among the population.

However, now that we found topics areas of interest of vaccine hesitancy, we think further data science projects should be conducted to discover more details of these causes, by possibly analysing tweets from just one of the most popular topics, like political reactions or opinions.

Contributions

Our team worked well together and was able to continuously exchange ideas and meet as soon as a problem appeared. We decided to be efficient and to put everyone's best qualities to use.

Massimo mainly focused on using the Twitter API and getting the tweets with the requirements we sought after in the data collection phase; Zachary worked on the TF-IDF analysis using the spaCy package, and Robert focused on the structure and analysis part of the report.

Of course, we were never alone in handling these parts, if needed we asked each other for support when a part of our tasks became too difficult to handle alone, often this was the case. We were all present for the initial discussion, where we decided what keywords to include and what filters to place on our tweets. We also came together for the data annotation phase, working together to make sure our topics and personal bias were in check. And of course, when all the data was collected, we came together for a meeting and discussed our results and findings, to try and put them into a cohesive argument. Despite exams and busy schedules, we were able to keep in touch, keeping everyone informed of any progress we made.