Opinion Mining on Coronavirus Vaccine Hesitancy

Saad Teeti
Computer Science, NYUAD
sht318@nyu.edu

Advised by: Prof. Mai Oudah

ABSTRACT

Vaccine Hesitancy is defined as delay in acceptance or refusal of vaccination despite the availability of vaccination services. The World Health Organization classified vaccine hesitancy as one of the ten threats to global health in 2019 which was even before the Covid-19 pandemic. In 2021 with the Covid-19 outbreak around the world, vaccine hesitancy is gradually becoming a larger and more dangerous issue that the health sector is facing around the world. Therefore, in this capstone project, we are researching more regarding this topic through using sentiment analysis in order to understand trends in such phenomena.

The research questions that we drafted to guide us through the process were; can we develop a Sentiment Analysis model that would successfully allow us to capture trends and potential reasons on the hesitancy of people over taking the coronavirus vaccine, which Machine Learning (ML) algorithm would work best for this problem, and potentially analyze the hesitancy over different types of vaccines with high accuracy via ML-based sentiment analysis? Can we visualize such trends?

Through our progress we analyzed and performed a literature review on previous work that had a similar idea to our topic but with different types of vaccinations, in addition we used these papers as inspiration to develop our experience and methodology. This included studying other research papers' approaches for building the machine learning models, approach on data retrieval, and visualizing trends that were founded. Throughout our research project we trained and built multiple Machine Learning (ML) based models to analyze tweets using Natural Language Processing (NLP) sentiment analysis. In addition we constructed a dataset that included tweets related to Covid-19 vaccines in general and Covid-19 vaccines produced by different companies as well. Using the model that we built, we captured trends, reasons and sentiments on the hesitancy of the people over taking the coronavirus vaccine.

KEYWORDS

Twitter, Natural Language Processing, Machine Learning, Vaccines Hesitancy, Coronavirus, Covid-19, Sentiment Analysis, Opinion Mining.

Reference Format:
Saad Teeti. 2022. Opinion Mining on Coronavirus Vaccine Hesitancy. In *NYUAD Capstone Project 2 Reports, Spring 2022, Abu Dhabi, UAE.* 13 pages.

1 INTRODUCTION

Vaccine Hesitancy is defined as the delay in acceptance or refusal of vaccination despite the availability of vaccination services [11]. The World Health Organization classified vaccine hesitancy as one of the ten threats to global health in 2019[8]. With the ongoing epidemic of Coronavirus that started towards the end of 2019, the hesitancy for people to take the Covid-19 vaccines is on an increase on a daily basis. Vaccines are considered one of the most successful public health interventions as vaccination led to the elimination and control of a lot of diseases in the past when people reached a point to accept such vaccines.

The refusal of taking vaccines can be caused by various reasons such as some peopling believing the need to reach herd immunity naturally without vaccination[5], questioning the safety of vaccines due to the short period time that they were developed, and the fear of potential side effects caused by vaccines [12]. Over time vaccines proved to be effective against different types of diseases such as measles, mumps, chickenpox, rubella, or tuberculosis by preventing millions of deaths globally every year [9].

In this project, our research questions are "Can we develop

This report is submitted to NYUAD's capstone repository in fulfillment of NYUAD's Computer Science major graduation requirements.

جامعة نيويورك أبوظبي
NYU ABU DHABI

Capstone Project 2, Spring 2022, Abu Dhabi, UAE

an Sentiment Analysis model that would successfully allow us to capture trends on the hesitancy of the people over taking the coronavirus vaccine, which ML algorithm would work best for this problem, and potentially analyze the hesitancy over different types of vaccines with high accuracy via ML-based sentiment analysis? Can we visualize such trends?"

2 RELATED WORK

Vaccine Hesitancy has always been a topic of interest in the research community as evident by the number of existing papers trying to study this topic and such trends, with Twitter frequently used as an environment for data gathering. The papers that have been published in this field that analyze Covid-19 vaccine hesitation, in particular, are limited because this is still considered a relatively new virus. An interesting paper that we came across was *Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy* [17]. According to this research paper, the motivation was a result of the drop in vaccination coverage, and the subsequent measles epidemic in 2017 with about 4885 cases and 4 deaths, which has attracted the interest of concerned experts, people, and the media, causing a heated political debate regarding vaccination in Italy. The purpose of this study was to find out and monitor public health opinion on vaccination concerning significant vaccine-related events that occurred at the time frame of September 2016 to August 2017. The study used three main methods that included the selection of tweets and preprocessing them, supervised learning stage and classification model accuracy, and data analysis towards the end. This paper was particularly helpful due to the emphasis placed on efficient visualization.

The research paper, *Vaccine Hesitancy on SocialMedia: Sentiment Analysis from June 2011 to April 2019* [16], has tackled vaccine hesitancy with the use of Twitter as an opinion mining tool. The techniques that they used in this paper were focused on the data source and data extraction where they used Twitter API to collect data using one keyword over a short period of time. From this data collection, 27 hashtags were identified and then using those 27 hashtags, the retrospective search of all public tweets was gathered (1,652,156 tweets excluding retweets). Then they prepared the data and did sentiment polarity analysis. In addition, they provided a statistical analysis using IBM Statistical Package for the Social Sciences where they evaluated trends by fitting linear, exponential, quadratic, cubic, and logistic models to the data and selecting the best fitting curve.

Optimization on machine learning-based approaches for sentiment analysis on HPV vaccines related tweets [4] has addressed the low intake of Human papillomavirus (HPV) vaccines. Throughout this paper, they analyze public opinions on HPV vaccines on social media using machine learning-based approaches to understand the reasons behind the low vaccine coverage and come up with corresponding strategies to improve vaccine uptake. The authors are proposing a machine learning system that can extract comprehensive public sentiment on HPV vaccines on Twitter with satisfying performance. Their methodology consisted of data collection, annotation schema design, tweets preprocessing and feature extraction, and building models using different machine learning algorithms where support vector machines performed the best. As a result, the researchers managed to provide a systematic way to improve machine learning model performance on highly unbalanced HPV vaccine-related tweets.

The work of Kuroshima and Tian [10] constitutes another interesting study in the medical domain. Even though it does not tackle vaccine hesitation directly, it focuses on people's sentiment on over-the-counter (OTC) drugs. This paper was especially helpful in describing a computational method that identifies the public sentiment towards over-the-counter (OTC) drugs using Twitter, which is a similar approach to what we are looking for. Their methodological approach begins by narrowing down the research to four OTC painkillers and a shorter list of symptoms. The two lists were passed through the Twitter API and retrieved tweets based on identified keywords. We will be following a similar approach in which we will be narrowing down our Twitter research to specific vaccine names. The other parts of their methodological approach included preprocessing the data and building machine learning models using decision tree learning, random forest, support vector machine, naïve Bayes, and k-nearest neighbors. The study uses the highest precision for automatic sentiment classification.

Another interesting research paper that we came across was *Comparing covariation among vaccine hesitancy and broader beliefs within Twitter and survey data* [14]. The research objective of this paper was to examine whether scientists can draw similar conclusions from Twitter and national survey data about the relationship between vaccine hesitancy and a broader set of beliefs. In 2018, these scientists conducted a nationally representative survey of parents in the United States informed by a literature review to ask their views on a range of topics, including vaccine side effects, conspiracy theories, and understanding of science. I liked the methodology used in this paper as they developed a set of keyword-based queries corresponding to each of the belief items from a survey they performed and pulled matching tweets from 2017.

The previous five papers discussed the practical approach of how they dealt with sentiment analysis using Twitter, however, Gohil et al. [7] discussed the methods used in a hypothetical approach. According to this paper, sentiment is a metric commonly used to investigate the positive or negative opinions within these messages. Exploring the methods used for sentiment analysis in Twitter-based healthcare research may allow us to better understand the options available for future research in this growing field. The study aimed to review the methods used to measure sentiment for Twitter-based health care studies. The first objective was to review what methods of sentiment analysis have been used and in which health care setting. The second objective was to explore to what extent the methods were trained and validated for the study data, and if any justification for their methodology use was offered. Their first method was identification and screening: where they narrowed down studies related to this field and their second part of the methodology was to compare the methods in each study. They looked at the methods of tools productions, in which setting they were used, and the method of testing these tools. For the assessment part of the paper, a comparison of the number of annotators was used to manually annotate tweets.

The previous papers focused on vaccine hesitancy before the coronavirus pandemic, however with time progressing, more papers regarding Covid-19 Hesitancy were published. In a research paper titled, *COVID-19 Vaccine Hesitancy in Canada: Content Analysis of Tweets Using the Theoretical Domains Framework*[8], the goal was to identify the main types of vaccine-hesitant tweeters. Their methodology included content analysis of vaccine-hesitant tweets, categorization of vaccine refusing tweeters, and data analysis of the results. In their results, they were able to find that the three major vaccine hesitation topics on Twitter accounted for half (50.2%) of the 446 tweets manually categorized: conspiracies (23.5%), development speed (16.1%) and safety (10.5%). However, the tweets were mostly from the USA and Canada, reflecting this national perspective in these countries.

Another study that was published recently, *COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies* [13], addressed anti-vaccine sentiment. In this paper, they presented a dataset of tweets that exhibit a strong anti-vaccine stance. The dataset consists of two parts: a) a streaming keyword-centered data collection with more than 1.8 million tweets, and b) a historical account-level collection with more than 135 million tweets. The former leverages the Twitter streaming API to follow a set of specific vaccine-related keywords starting from mid-October 2020. The latter consists of all historical tweets of 70K accounts that were engaged in the active spreading of anti-vaccine narratives. In addition, they presented descriptive analyses showing the volume of activity over time, geographical distributions, topics, news sources, and inferred accounts of political leaning. This dataset can be used in studying anti-vaccine misinformation on social media and enable a better understanding of vaccine hesitancy.

3 SPECIFIC AIMS AND OBJECTIVES

Understanding why people are refusing to take the Coronavirus vaccine and trying to measure trends of how people are open to taking the vaccines could be an important aspect in eventually eliminating the virus and achieving immunity. A social media platform to understand people's opinions on a topic could be Twitter. Using Twitter, we could do opinion mining where we try to trace, sentimentally classify, and find trends that are related to why people are not taking such vaccinations. Visualizing such trends by analyzing the sentiment of tweets related to vaccination could help the science community to overcome such obstacles and find other solutions to convince people to take vaccines.

4 DATASETS

The datasets used in this project can be categorized as follows:

- **Annotated Dataset:** An existing dataset that consists of 1.6 million tweets (23,011,409 words) including their sentiment, which we used to train and test the ML classifiers. The annotated dataset is from Kaggle, Go [6]'s research work, in particular. The dataset was made up of 50% of negative sentiment tweets and 50% of positive sentiment tweets. We split the dataset into 80% training set and 20% testing set. Using this dataset, we were able to create a binary classifier that would help us annotate tweets regarding vaccine hesitation.
- **Covid-19 Vaccine-Related Tweets:** We constructed a dataset that consists of 628,582 tweets to study the hesitancy of taking Covid-19 Vaccines. Our approach was through identifying specific keywords and having a controlled experiment where we input specific Twitter queries and retrieve tweets accordingly. We have used premium Twitter API for Academic Research, which gave us access to the full archive of tweets, allowing us to go back to any period of time and retrieve tweets from specific time intervals. We further elaborate on our approach to how we constructed this dataset in the coming section.

5 METHODOLOGY

Throughout this section, we will be talking about the methodology we followed in building sentiment analysis models, in addition to constructing the Covid-19 vaccine-related dataset.

5.1 Building Sentiment Analysis Models

5.1.1 Importing the Dataset. The already annotated dataset that we used is the sentiment140 dataset, where tweets have been annotated to 0 for negative sentiment and 4 for positive sentiment. From the dataset, we had 6 fields that we could work with, which are sentiments, ids, data, flag, user, and text. However, we only used the text and the sentiment to train the models as the other aspects are considered unnecessary to build a sentiment classifier.

5.1.2 Preprocessing the Text. To get the pure sentiment of the sentence, we had to preprocess the text by replacing full-length URLs with the token "URL", emojis with their value (i.e smiling face to "happy"), and usernames with "USER". In addition, we removed non-alphabets, consecutive letters, short words, and stop words. In addition, we also performed lemmatization and lower casing on the dataset.

5.1.3 Analyzing the Data. After preprocessing the text, we wanted to analyze the data by visualizing the frequency for specific words representing negative and positive sentiment. The purpose of analyzing the data and finding the most repetitive words was to identify which words are mostly used for each sentiment. In Figures 1 and 2, we can see that the word "USER" was heavily repeated in both positive and negative tweets, and the only reason for that is that we replace all the usernames in every tweet with the word "USER".

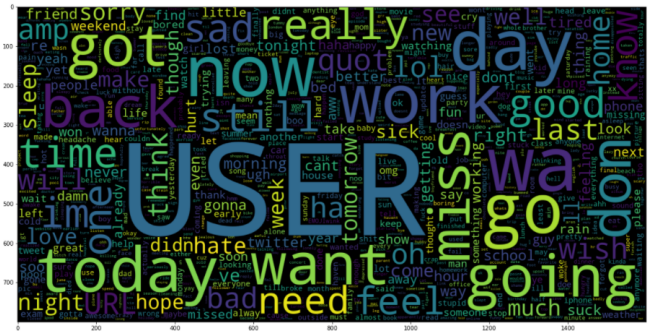

Figure 1: A word cloud representation of the most used words for the negative sentiment.

Figure 2: A word cloud representation of the most used words for the positive sentiment.

5.1.4 Splitting the Dataset. After preprocessing the data, we split it into two datasets, including 80% for training and 20% for testing.

5.1.5 TF-IDF Vectorization and Transforming the dataset. TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency [1]. To build our sentiment classification model, we need to extract features to produce a feature set of vectors representing every word in the dataset. We used TF-IDF vectorization which indicates the importance of the word, and that is considered necessary to understand the dataset. Using this vectorization method, the first feature we extracted was the term frequency, which is the occurrence of the word in the dataset over the total number of words in the dataset. The second feature that was extracted was the weight of each word based on its frequency that occurred in the dataset. Using TF-IDF built-in function through SK-learn python libraries [15], we were able to extract features, fit and transform the data using the vectorizer, and use it to build the models.

5.1.6 Creating Sentiment Classifiers. After following the previous steps, we were ready to train our models using the extracted feature set. We explored the performance of three machine learning (ML) algorithms chosen based on the conducted literature review, which identifies the best performing ML algorithms. The first model that we built using the Naive Bayes Classification algorithm [18], the second model that we built using the Support Vector Classification algorithm [2], and for the third built model, we used Logistic Regression [3]. After training the three models, we evaluated them based on the precision, recall, and f1 score, which we will discuss in the next subsection.

5.2 Models Results and Evaluation

Overall the evaluation results from the three models were relatively close. We present and discuss the results of each of

	precision	recall	f1-score	support
0	0.81	0.79	0.80	159815
1	0.80	0.81	0.80	160185
accuracy			0.80	320000
macro avg	0.80	0.80	0.80	320000
weighted avg	0.80	0.80	0.80	320000

Table 1: Evaluation Table for Naive Bayes Classification Model

	precision	recall	f1-score	support
0	0.82	0.81	0.81	159815
1	0.81	0.82	0.82	160185
accuracy			0.82	320000
macro avg	0.82	0.82	0.82	320000
weighted avg	0.82	0.82	0.82	320000

Table 2: Evaluation Table for Support Vector Classification Model

the three models in this section. For our experimental setting, we built a pipeline to evaluate our model. In this pipeline, we passed the models that we trained with the fitted X and Y training sets, which we split previously from the dataset, (80% of the original dataset). We used the models that we passed to predict outcomes/labels for the test dataset (20% of the original dataset). Then for results visualization, we used the heatmap of confusion matrix to display the evaluation.

5.2.1 Naive Bayes Classification Model. Using the built-in Naive Bayes classifier provided with SK-Learn Library, we built the model, fitted the X and Y training datasets, and then passed the model through the evaluation function. Overall our accuracy was 80% for this model. Refer to Table 1 for the precision, recall, f1-score, and support results (label 0 refers to positive labels and label 1 refers to negative labels). Refer to Figure 3 for the confusion matrix results.

5.2.2 Support Vector Classification Model. Using the built-in Support Vector Classifier provided with SK-Learn Library, we built the model, fitted the X and Y training datasets, and then passed the model through the evaluation function. Overall our accuracy was 82% for this model. Refer to Table 2 for the precision, recall, f1-score, and support results (label 0 refers to positive labels and label 1 refers to negative labels). Refer to Figure 4 for the confusion matrix results.

5.2.3 Logistic Regression Model. Using the built-in Logistic Regression Model provided with SK-Learn Library, we first

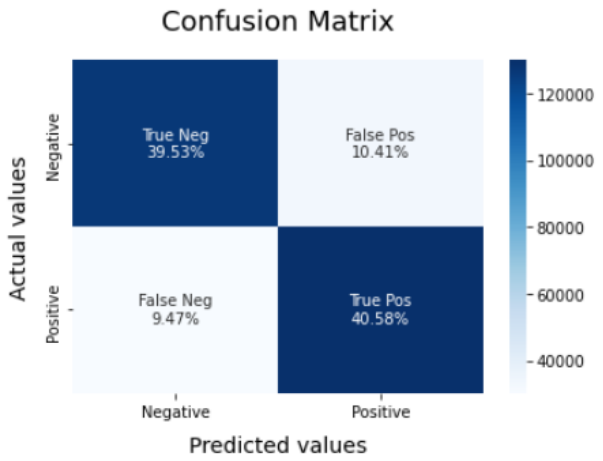

Figure 3: Confusion Matrix representation for Naive Bayes Classification Model

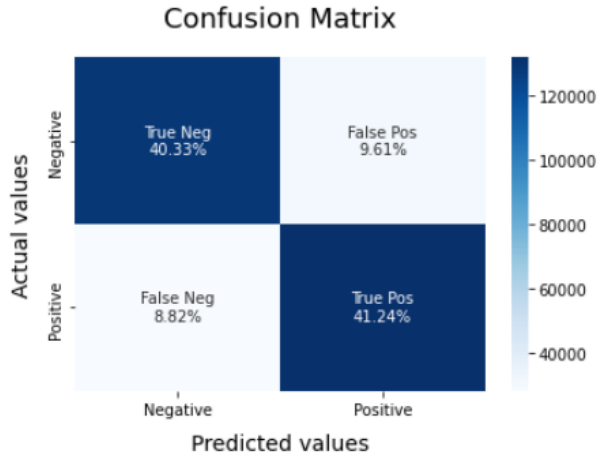

Figure 4: Confusion Matrix representation for Support Vector Classification Model

set the max iteration for the model to be 1000, then we fitted the X and Y train datasets, and then we passed it through the evaluate function. Overall our accuracy was 83% for this model. Refer to Table 3 for the precision, recall, f1-score, and support results(label 0 refers to positive labels and label 1 refers to negative labels). Refer to Figure 5 for the confusion matrix results.

	precision	recall	f1-score	support
0	0.83	0.82	0.82	159815
1	0.82	0.84	0.83	160185
accuracy			0.83	320000
macro avg	0.83	0.83	0.83	320000
weighted avg	0.83	0.83	0.83	320000

Table 3: Evaluation Table for Logistic Regression Model

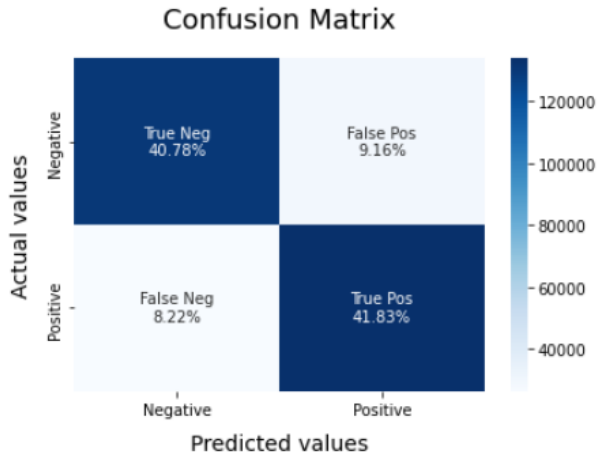

Figure 5: Confusion Matrix representation for Logistic Regression Model

5.3 Retrieving Covid-19 Vaccine-Related Tweets

To construct our datasets, we created a pipeline to retrieve tweets regarding the Covid-19 vaccination. Using the Academic Twitter API, we had access to both full archive tweets and full-length tweets. To make sure that we are retrieving relevant tweets, we started by testing different queries and time frames that we want to use for data extraction. Among setting up our queries we tested different formats of queries and examined the data returned to us, and by then we optimized our queries to extract relevant data for our research project.

The first parameter that we were testing was the time frame at which we wanted the retrieval process to start and after looking at different tweets from different time frames, we decided that it would be best if we retrieve tweets starting from the date of December 2020, where Covid-19 vaccination trails just started and before they were issued for the public.

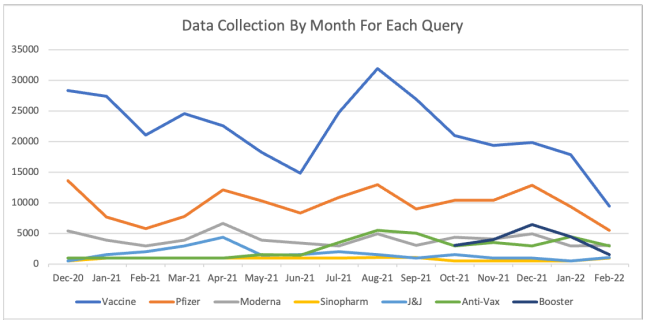

Figure 6: Data Collection By Month For Each Query

Our datasets consist of tweets from December 1st, 2020 to February 28, 2022. The second parameter was the language of the tweet, where we chose only English language tweets. The third parameter was excluding tweets that were classified as ads. The fourth parameter included the keywords that needed to be matched in the tweets for retrieval, which were split into 7 different categories. So overall we had 7 different queries that we used to retrieve the data that included the time frame, language, excluding ads features, and keywords.

Figure 6 shows the numbers of tweets collected per month for each query. The first 6 queries consisted of the same time frame, language, excluding ads feature, but different keywords. The seventh query time frame was different from the other six queries. We will further explain the details for each query used in the subsections bellow:

(1) The methodology that we used for the first query was to retrieve tweets about covid vaccination in general, excluding companies that make the vaccine. The keyword search query was constructed of three conditions. The first condition was to include either the following keywords (vaccine or vaccines or vaccination), the second part was to include (Covid-19, covid, corona, or coronavirus). The third part was to exclude companies that manufacture the vaccine and other names of common vaccines. The following sets of keywords were used to exclusion (Johnson & Johnson, Pfizer, Moderna, Sinopharm, booster, Adenovirus, Anthrax, Cholera, Diphtheria, Hepatitis A, Hepatitis B, Hepatitis, HPV, FLU, Measles, Meningococcal, Mumps, Pertussis, Pneumococcal, Polio, rabies, rotavirus, rubella, shingles, smallpox, tetanus, Tuberculosis, Typhoid Fever, Varicella, and Yellow Fever). Overall from this query, we retrieved 327,991 over 15 months.
(2) The second query was targeting tweets related to vaccines produced by Pfizer, an American pharmaceutical company. The search query consisted of the keyword

"Pfizer". Overall from this query, we retrieved 146,755 over 15 months.

(3) The third query was targeting tweets related to vaccines produced by Moderna, an American pharmaceutical company. The search query consisted of the keyword "Moderna". Overall from this query, we retrieved 60,222 over 15 months.
(4) The fourth query was targeting tweets related to vaccines produced by Sinopharm, a Chinese pharmaceutical company. The search query consisted of the keyword "Sinopharm" or "Sinovax". Overall from this query, we retrieved 12,315 over 15 months.
(5) The fifth query was targeting tweets related to vaccines produced by Johnson and Johnson, an American pharmaceutical company. The search query consisted of the keywords "johnson and johnson" or, "Johnson-Johnson" or "Johnson Johnson". Overall from this query, we retrieved 23,626 over 15 months.
(6) By the sixth query, we wanted to retrieve tweets related to anti-vax directly. The search query included keywords "antivax" or "anti-vax" or "anti vax". Overall from this query, we retrieved 38,343 over 15 months.
(7) By the last query, we wanted to retrieve and examine opinions regarding the Booster shot. Booster shots became relevant later on in the vaccination process. We decided to have a time frame of 5 months from October 2021 to February 2021. The search query included two parts, first the keyword Booster, and the second part was "shot" "shots". Overall from this query, we retrieved 19,330 over 5 months.

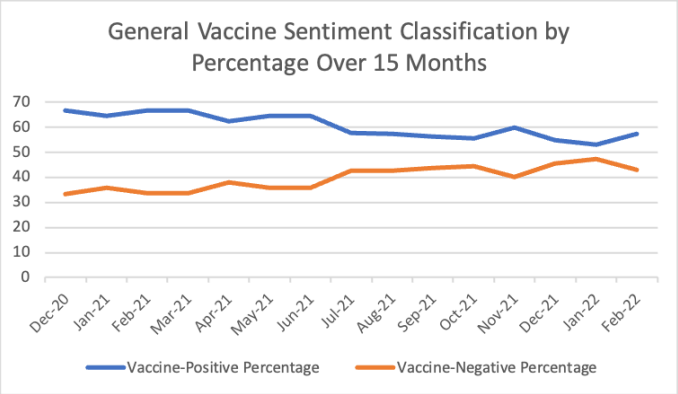

Figure 7: General Vaccine Sentiment Classification by Percentage Over 15 Months

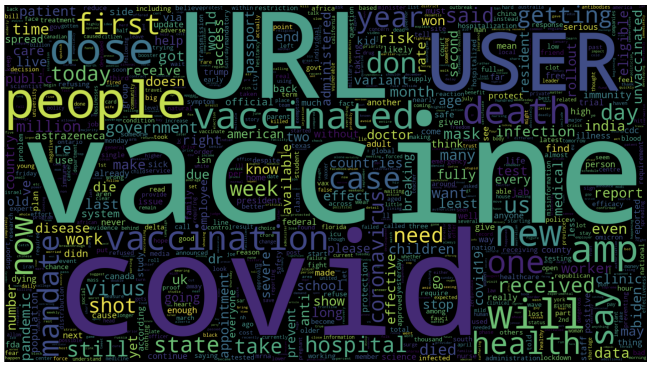

Figure 8: Word Cloud for Negative General Vaccine Sentiment

6 APPLYING THE MODEL

Since Logistic Regression Model yielded the highest accuracy, we used it to annotate the retrieved tweets. Using the same method used in 5.1.2 to preprocess the text, we preprocessed the tweets that we retrieved. After we preprocessed the text we used the vectorizer that we built in section function 5.15 to extract features and then we applied the logistic regression model for each set of tweets retrieved based on the queries.

6.1 General Vaccine Related Tweets

The first query which included general Covid-19 vaccine-related tweets yielded results of 61% positive classified tweets and 39% negative classified tweets. Overall 198,732 tweets were labeled as positives while 129,259 tweets were labeled as negatives. Figure 7 shows the tweets classification by months, and figure 8 shows the word cloud for negative labeled tweets.

6.2 Pfizer Vaccine Related Tweets

The second query which included Pfizer Covid-19 vaccine-related tweets yielded results of 56% positive classified tweets and 44% negative classified tweets. Overall 82,721 tweets were labeled as positives while 64,034 tweets were labeled as negatives. Figure 9 shows the tweets classification by months, and figure 10 shows the word cloud for negative labeled tweets.

6.3 Moderna Vaccine Related Tweets

The third query which included Moderna Covid-19 vaccine-related tweets yielded results of 60% positive classified tweets and 40% negative classified tweets. Overall 36,334 tweets were labeled as positives while 23,888 tweets were labeled as negatives. Figure 11 shows the tweets classification by months, and figure 12 shows the word cloud for negative labeled tweets.

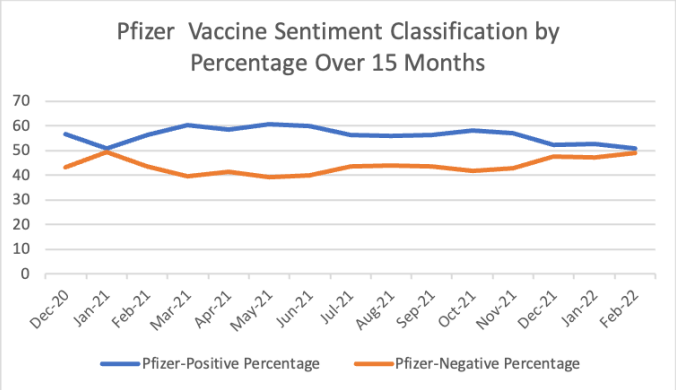

Figure 9: Pfizer Vaccine Sentiment Classification by Percentage Over 15 Months

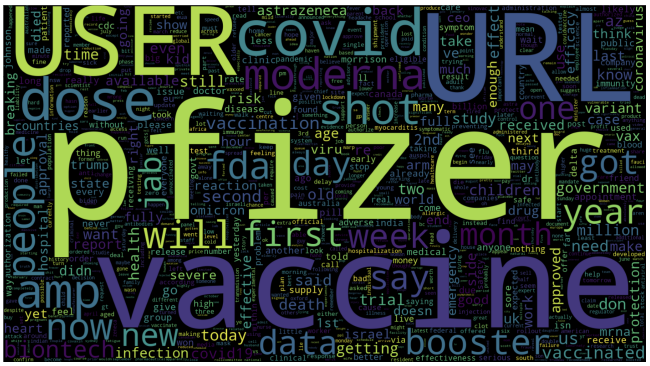

Figure 10: Word Cloud for Negative Pfizer Vaccine Sentiment

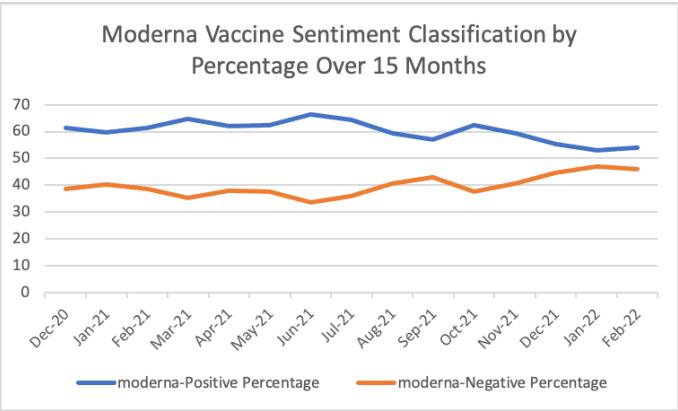

Figure 11: Moderna Sentiment Classification by Percentage Over 15 Months

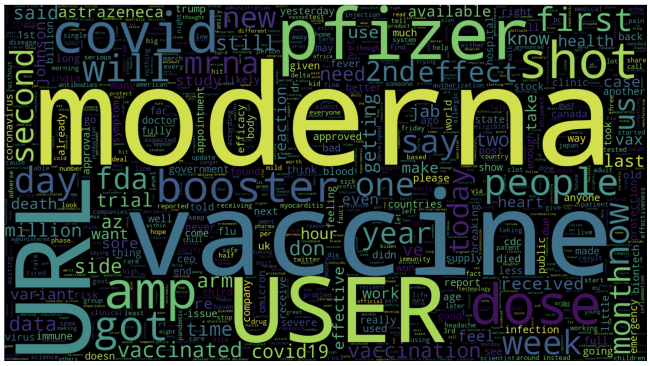

Figure 12: Word Cloud for Negative Moderna Vaccine Sentiment

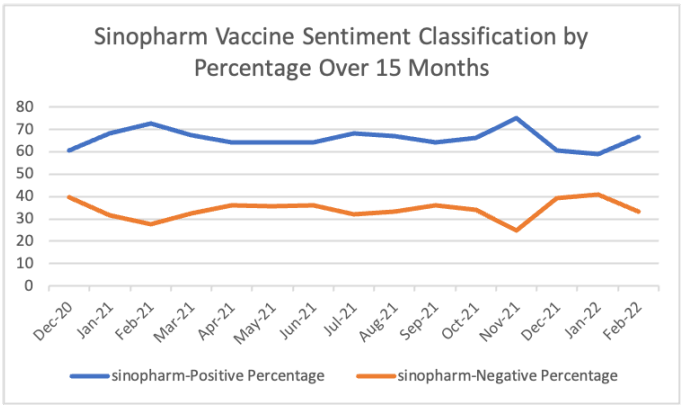

Figure 13: Sinopharm Sentiment Classification by Percentage Over 15 Months

6.4 Sinopharm Vaccine Related Tweets

The fourth query which included Sinopharm Covid-19 vaccine-related tweets yielded results of 66% positive classified tweets and 34% negative classified tweets. Overall 8,146 tweets were labeled as positives while 4,169 tweets were labeled as negatives. Figure 13 shows the tweets classification by months, and figure 14 shows the word cloud for negative labeled tweets.

6.5 Johnson&Johnson Vaccine Related Tweets

The fifth query which included Johnson & Johnson Covid-19 vaccine-related tweets yielded results of 77% positive classified tweets and 23% negative classified tweets. Overall 18,187 tweets were labeled as positives while 5,439 tweets were labeled as negatives. Figure 15 shows the tweets classification by months, and figure 16 shows the word cloud for negative labeled tweets.

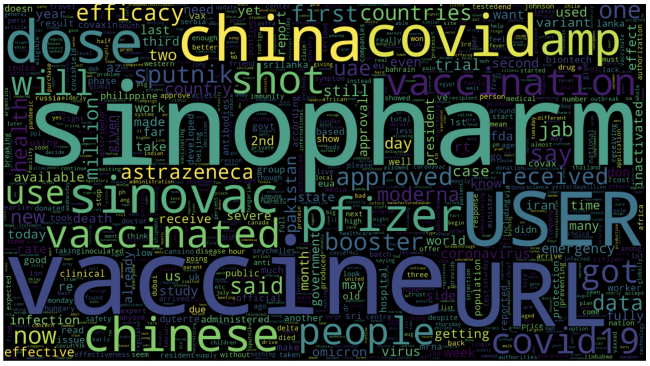

Figure 14: Word Cloud for Negative Sinopharm Vaccine Sentiment

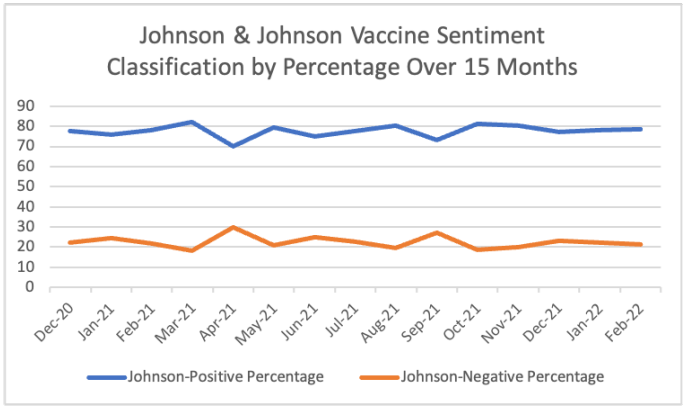

Figure 15: Johnson & Johnson Sentiment Classification by Percentage Over 15 Months

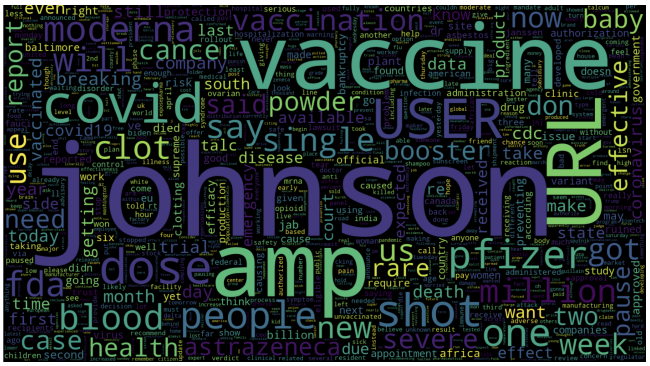

Figure 16: Word Cloud for Negative Johnson & Johnson Vaccine Sentiment

6.6 Anti-vax Vaccine Related Tweets

The sixth query which included Anti-Vax Covid-19 vaccine-related tweets yielded results of 41% positive classified tweets and 59% negative classified tweets. Overall 15,777 tweets were labeled as positives while 22,566 tweets were labeled as negatives. Figure 17 shows the tweets classification by months, and figure 18 shows the word cloud for negative labeled tweets.

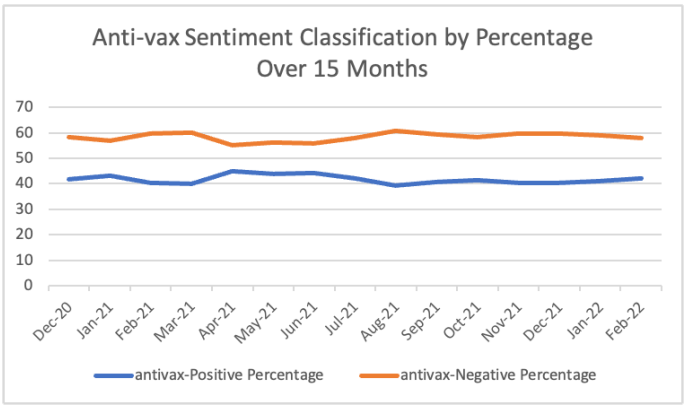

Figure 17: Anti-vax Sentiment Classification by Percentage Over 15 Months

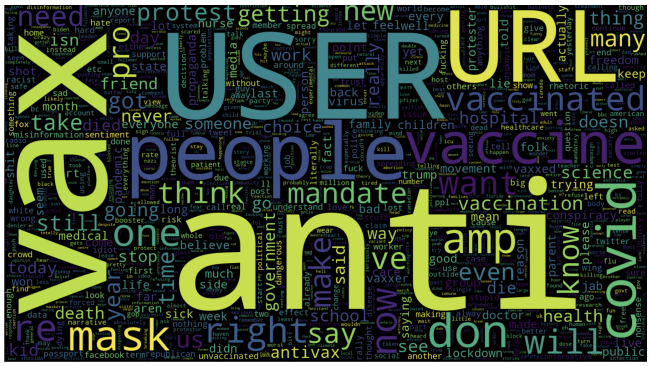

Figure 18: Word Cloud for Negative Anti-vax Vaccine Sentiment

6.7 Booster Vaccine Related Tweets

The seventh query which included Booster Covid-19 vaccine-related tweets yielded results of 55% positive classified tweets and 45% negative classified tweets. Overall 10,661 tweets were labeled as positives while 8,669 tweets were labeled as negatives. Figure 19 shows the tweets classification by months, and figure 20 shows the word cloud for negative labeled tweets.

6.8 Sentiment Misclassification

During the process, some tweets were misclassified by the model. The reasoning for tweets misclassification could vary, but we tried to find out some of the reasons why by analyzing

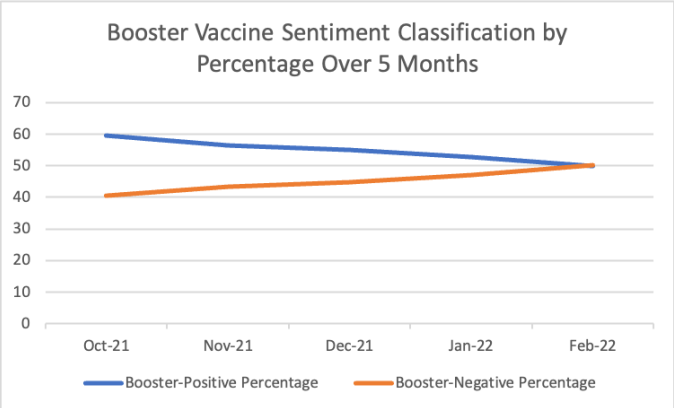

Figure 19: Booster Sentiment Classification by Percentage Over 15 Months

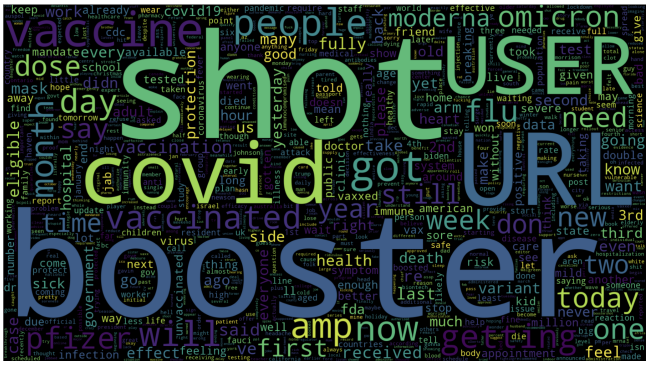

Figure 20: Word Cloud for Negative Booster Vaccine Sentiment

samples of the misclassified tweets. Our analysis concluded three main reasons.

6.8.1 Sarcasm, irony, and humor. Sarcasm, irony, or humor make it harder for the model to identify sentiment properly. An example of a tweet that included humor and was misclassified as positive:

- "USER awesome so guys who pretend to be physicians while they spend their days doing donald bidding get vaccines while front line healthcare workers who battle covid 19 everyday are still waiting not good look greg"

6.8.2 Negations. Another potential reason why the model could misclassify tweets is negation terms. For example, "not bad" means good, however the model may classify it as negative because of the token "bad". The example below shows how one of the tweets was misclassified as negative due to negation:

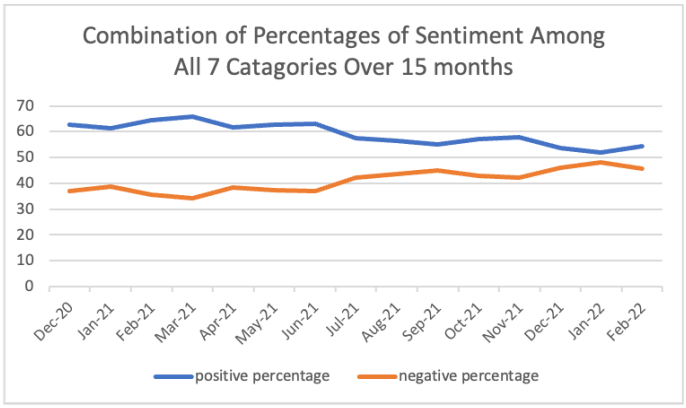

Figure 21: Combination of Percentages of Sentiment Among All 7 Categories Over 15 months

- "This week was not bad for us! Our covid-19 cases are going down. Many ugandans have still not accessed the vaccine. We are also still under lockdown In Uganda. Most of the work we are doing is virtual SheLeads Nyayadzedu."

6.8.3 Multi-polarity. Multi-polarity tweets might also get misclassified as they would include more than one keyword that would have the opposite polarity. The examples below show tweets that included two different polarities. The first tweet was misclassified as negative and the second tweet was misclassified as positive:

- "will say the one good thing about me and nick getting covid and him getting really sick is that it convinced my granny to get the vaccine and think if he didn get sick she still would be unvaccinated "
- "just recovered from pretty bad case of covid and if given the choice now between covid and the vaccine would choose covid again "

7 ANALYSIS AND DISCUSSION

7.1 Monthly Trends

By counting all the annotations among different categories as shown in Figure 21, we can see that the negative sentiment rate related to COVID-19 vaccine on twitter increased from March 2021 and onward. Between February and March of 2021, vaccines started being produced more and were offered to the public after passing the trails, and hence why the negative sentiment started increasing as vaccines were becoming more relevant and involved in COVID-related protocols.

7.2 Negative Sentiment Reasons Extraction

Using the negatively classified tweets, we tried to extract the most frequent and common used words to infer reasons

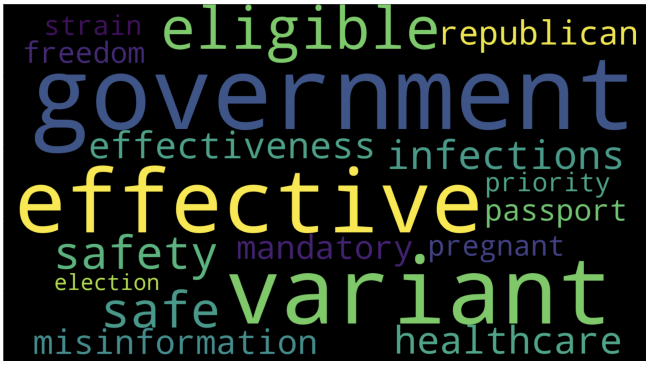

Figure 22: Word Cloud for Most Frequent, Common and Relevant Words Found in Negatively Classified Tweets Across All Categories

why people are hesitant to take the vaccine. Figure 22 shows word cloud most frequent keywords found in the negatively classified tweets across all categories. Below we list some of the most frequent keywords that appeared in the negative tweets in an attempt to provide reasons behind vaccine hesitancy, along with examples for each indicator in terms of tweets.

7.2.1 ***Government:*** Concerns and criticism over government responses to Covid-19 out-break and how they handled the situation when it came to providing and manufacturing vaccines.

- "turned it off heard absolutely nothing that makes me worried or concerned except possibly the government response the vaccine data shows it works against variants the amount of past covid infections amp vaccine data means outbreak would be near impossible enjoy weekend"

7.2.2 ***Effectiveness, Strains, and Variants:*** Concerns over the effectiveness of the vaccine against new variants and strains.

- "South Africa has just rejected the AstraZeneca vaccine against COVID-19. It is not effective for the South African strain of the virus. Why should the rest of Africa receive millions of doses when virus strains are undetermined?"
- "Just In:- .New strain of Covid-19 virus found in England. Vaccine seems to be not effective"
- "My 2 worst fears as an ER doc 1) New variants of COVID-19 not protected by the vaccine 2) New variants that will kill infants amp; children disproportionately You know how we can overcome both? 1) Get vaccinated, if vaccines are available to you. 2) Make vaccines available"

7.2.3 ***Safety:*** Concerns over the safety of the vaccine.

- "Why is our government pushing a partially tested vaccine which may have dangerous side effects versus using a safe, cheap, and extremely effective prophylaxis and treatment for covid-19 that is ivermectin? "
- Bobby Kennedy's concern â that the Covid vaccine is potentially unsafe, and has not been properly tested â is widespread, and dangerously wrong,â writes @drkerrymeltzer, his niece, and a physician.

7.2.4 ***Priority and Eligibility:*** Concerns over the priority and the eligibility of the vaccination drive.

- "Updated statement from the JCVI: unvaccinated adults aged 30 to 39 years who are not in a clinical priority group at higher risk of severe COVID-19 disease, should be preferentially offered an alternative to the AstraZeneca COVID-19 (AZD1222) vaccine"

7.2.5 ***Infections:*** Concerns over vaccines not preventing further infections.

- "the covid cultists wants it both ways with the vaccines they say they are effective amp everyone has to get one amp they want vaccine passports but we can never ever go back to normal because the vaccines don prevent infections or transmission they only lessen symptoms"

7.2.6 ***Healthcare:*** Concerns over the healthcare system and people's lack of trust in such a system in the United States, in particular.

- "With 34% of South Dakotans having received at least one dose of a COVID-19 vaccine, the state's healthcare systems are left with thousands of tablets of hydroxychloroquine that former President Donald Trump claimed to be effective in preventing the"
- "The speed at which we got a vaccine for COVID-19 is what a quality market-driven health care system could do for America. "
- "Despite pervasive warnings about a lack of healthcare access and heightened vaccine hesitancy, Americas COVID-19 vaccine drive is failing to reach Black and Hispanic communities"

7.2.7 ***Misinformation:*** Misinformation regarding lies and conspiracies that are being spread around.

- "One of the biggest barriers standing in the way of ending the pandemic is not medical or logistical. It's the misinformation about the COVID-19 vaccines. "
- "Dear experts- doctors, scientists, politicians, MSM, etc., who are censoring covid and covid vaccine "misinformation", Only liars and cheats feel the need to censor others. The only reason to censor and silence others is because your science can't hold up to scrutiny."

- "People are dying like flies and the govt is busy with sheninegans to descredit the existence of covid 19. The president is busy with misinformation and theories on vaccines and existance of covid 19 which is purely against what science says JK and your team you take the BLAMES"

7.2.8 ***Republicans:*** Concerns over republicans' refusal to take the vaccines.

- "rachel maddow account of the dire need many countries have to get covid vaccines and how the Biden administration wants to break patents so these countries can get produce them who is opposed to that republican they are really sociopaths it global health crisis"

7.2.9 ***Mandatory, freedom and Covid Passport:*** Freedom restriction and the inability to travel without a vaccination passport.

- "Can anyone tell me do you think we will have COVID passports, I am against the untested vaccine and this news by blare and piers Morgan is so depressing, is there any hope left, I dont want to take this vaccine and will fight it, but have we lost the fight? So down right now"
- "We must take note of every politician of any party who expresses support for mandatory covid vaccination or vaccine passports or digital credentials. They must never again be allowed to hold public office."
- "You should not be getting the Covid vaccine just so that your life can get back to normal. Medical decisions should be made based on having analyzed the pros/cons of the vaccine and not the threat of having your freedoms restricted."

7.2.10 ***Pregnancy:*** Concerns over vaccine safety when it comes to pregnant women.

- "Brazil has suspended the use of AstraZeneca's COVID-19 vaccine in pregnant women nationally after the death of an expectant mother. The coordinator of the Health Ministry's vaccination program said the suspension was enacted as a precautionary measure"
- "not sure how to say this forcefully enough without being uncharacteristically profane but to be very clear the covid 19 vaccines do not shed it utter bollocks this idea that vaccinated person can somehow harm pregnant woman this isn rocket science"

7.2.11 ***Election:*** Also, encouraging people to vote for a specific party that does not support mandatory vaccination.

- "Domestic vaccine passports won't be temporary and they won't stop at COVID. They are the thin end of the wedge. No Pass? No gym, no beer, no theatre... The elections tomorrow may be your last chance to be heard for a long time. Please use your vote wisely."

8 CONCLUSION AND FUTURE WORK

Vaccine Hesitancy is a very important topic that the social community has been struggling with recently. Throughout this project, we conducted opinion mining on Twitter as an approach to understanding and visualizing the trends of Coronavirus Vaccine Hesitancy. Throughout this research, we were successfully able to build 3 machine learning models that yielded a relatively high performance after testing them, and we were able to answer our research question on which machine learning model is the best fit for our research as Logistic Regression algorithm yielded the best performance. We were able to utilize both Natural Language Processing and Machine Learning to build these models. In addition, we were able to construct a dataset by retrieving tweets related to Covid-19 vaccines, which were necessary to analyze vaccine hesitancy.

We chose the classifier that produced the highest f1 score and we applied it to the dataset that we constructed. We analyzed hesitancy over time and we tried to elicit potential reasons behind why people are negative about the vaccine. Overall, there is still room for improvement that we could focus on in future work, such as enhancing the pipeline producing the classification models for better performance, and adding another category of classification that is Neutral to address this category of sentiment.

Another aspect that we would like to address in the future is the nature of tweets on the Twitter platform. There might be some propaganda within Twitter, and tweets that are coming through bots to increase a specific sentiment and enhance such propaganda.

REFERENCES

[1] Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. 2016. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications* 7, 4 (2016), 285–294.

[2] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[3] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232.

[4] Jingcheng Du, Jun Xu, Hsing-Yi Song, Xiangyu Liu, and Cui Tao. 2017. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics* 8 (03 2017). https://doi.org/10.1186/s13326-017-0120-6

[5] Alessio Facciolà, Giuseppa Visalli, Annalisa Orlando, Maria Paola Bertuccio, Pasquale Spataro, Raffaele Squeri, Isa Picerno, and Angela Di Pietro. 2019. Vaccine hesitancy: An overview on parents' opinions about vaccination and possible reasons of vaccine refusal. *Journal of*

public health research 8, 1 (2019).

[6] Alec Go. 2009. Sentiment Classification using Distant Supervision.

[7] Sunir Gohil, Sabine Vuik, and Ara Darzi. 2018. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. *JMIR Public Health Surveill* 4, 2 (23 Apr 2018), e43. https://doi.org/10.2196/publichealth.5789

[8] Janessa Griffith, Husayn Marani, and Helen Monkman. 2021. COVID-19 Vaccine Hesitancy in Canada: Content Analysis of Tweets Using the Theoretical Domains Framework. *J Med Internet Res* 23, 4 (13 Apr 2021), e26874. https://doi.org/10.2196/26874

[9] Inaya Hajj Hussein, Nour Chams, Sana Chams, Skye El Sayegh, Reina Badran, Mohamad Raad, Alice Gerges-Geagea, Angelo Leone, and Abdo Jurjus. 2015. Vaccines through centuries: major cornerstones of global health. *Frontiers in public health* 3 (2015), 269.

[10] Daisuke Kuroshima and Tina Tian. 2019. Detecting Public Sentiment of Medicine by Mining Twitter Data. *International Journal of Advanced Computer Science and Applications* 10, 10 (2019). https://doi.org/10.14569/IJACSA.2019.0101001

[11] Noni E MacDonald et al. 2015. Vaccine hesitancy: Definition, scope and determinants. *Vaccine* 33, 34 (2015), 4161–4164.

[12] Chephra McKee and Kristin Bohannon. 2016. Exploring the reasons behind parental refusal of vaccines. *The journal of pediatric pharmacology and therapeutics* 21, 2 (2016), 104–109.

[13] Goran Muric, Yusong Wu, and Emilio Ferrara. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR Public Health Surveill* 7, 11 (17 Nov 2021), e30642. https://doi.org/10.2196/30642

[14] Sarah A. Nowak, Christine Chen, Andrew M. Parker, Courtney A. Gidengil, and Luke J. Matthews. 2020. Comparing covariation among vaccine hesitancy and broader beliefs within Twitter and survey data. *PLOS ONE* 15, 10 (10 2020), 1–16. https://doi.org/10.1371/journal.pone.0239826

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[16] Hilary Piedrahita-Valdés, Diego Castillo, Javier Bermejo, Patricia Guillem-Saiz, Juan Bermejo-Higuera, Javier Guillem-Saiz, Juan Antonio Montalvo, and Francisco Machio. 2021. Vaccine Hesitancy on Social Media: Sentiment Analysis from June 2011 to April 2019. *Vaccines* 9 (01 2021), 28. https://doi.org/10.3390/vaccines9010028

[17] Lara Tavoschi, Filippo Quattrone, Eleonora D'Andrea, Pietro Ducange, Marco Vabanesi, Francesco Marcelloni, and Pier Lopalco. 2020. Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Human Vaccines Immunotherapeutics* 16 (03 2020), 1–8. https://doi.org/10.1080/21645515.2020.1714311

[18] Geoffrey I. Webb. 2010. *Naïve Bayes.* Springer US, Boston, MA, 713–714. https://doi.org/10.1007/978-0-387-30164-8_576