BRUNEL UNIVERSITY LONDON

CS5500 Dissertation 2022/3-Dissertation (12,000 words)

CS5500 Task 2 Dissertation 2022/3

Predefined Information

Start date:2023-09-02 09:00 AM BST

End date:2023-09-13 11:00 AM BST

Grading scale: British 17-point (A*-F)

Flow code:KT06928892

Internal assessor:(Anonymised)

Participant

Name:	SOORAJ KRISHNAKUMAR
Candidate No.:	2267641/1#CS5500_CB#2022/3#A#TRM2#CS5500_CB#001#P
Brunel id:	2267641@brunel.ac.uk
Brunel Student Id:	2267641

Department of Computer Science

MSc Data Science and Analytics

Academic Year 2022-2023

Exploring concerns on COVID-19 vaccine side effects on Reddit discussions

Sooraj Krishnakumar - 2267641

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

Brunel University
Department of Computer Science
Uxbridge, Middlesex UB8 3PH
United Kingdom
Tel: +44 (0) 1895 203397
Fax: +44 (0) 1895 251686

ABSTRACT

This dissertation investigates vaccination hesitancy, with an emphasis on the influence of social media and worries about myocarditis as a side effect of the COVID-19 vaccine. Understanding the factors that contribute to vaccine hesitancy is essential to developing health communication tactics that are effective. In light of the continuing COVID-19 pandemic, the study is especially pertinent.

In this work, Reddit postings were analysed using Natural Language Processing (NLP) and Machine Learning (ML) methods. Tokenization, lemmatization, and entity recognition are just a few of the NLP techniques that were used once the data had been cleansed and pre-processed. The topics mentioned in the posts were identified using topic modelling, which was done using Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), and the records that weren't relevant for the analysis were removed. The overall sentiment of the posts was also examined using sentiment analysis.

From this dissertation, we were able to deduce that the impact of myocarditis as a side effect of the COVID-19 vaccine is higher than that of many other side effects. Also, there is correlation between public concern about myocarditis and the reports from Government websites on the same. These conclusions were deduced using keyword analysis where Word clouds and collections were the packages used. Machine Learning algorithms were used to label the unlabelled data whereas Topics Modelling techniques such as LDA and NMF were used for removing the topics which are irrelevant to us in the analysis.

The key implication is that it analyses the data and provides us insights into the drivers of vaccine hesitancy and in the process able to introduce new policies to enhance vaccine confidence among the people. It also demonstrates that social media analysis can give us useful information not just in the medical field but in any sector.

.

ACKNOWLEDGEMENTS

I am grateful to have had you, Dr. Timothy Cribbin, as both Professor and Dissertation Supervisor. Your education and contribution helped me to set a clear direction to follow and inspired me to do better.

Furthermore, I offer my deepest gratitude to the Ethics department which provided me with suggestions as part of my dissertation project.

I would like to thank all for the wonderful time and thank each and every Professor and management who helped me during my course at Brunel University London.

I certify that the work presented in the dissertation is my own unless referenced

Signature...........Sooraj Krishnakumar

Date..........13/09/2023

TOTAL NUMBER OF WORDS: 8237

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The whole world was taken aback by the emergence of the COVID-19 pandemic, and it has been a daunting task to overcome the damage it has caused in many aspects of life since then. Since its start, the virus has wreaked havoc, with fatality cases and the number of people affected on a daily basis (Melton, 2022). The onset of the COVID-19 pandemic, caused by the new coronavirus SARS-CoV-2, quickly transformed the worldwide landscape, leaving a lasting effect on public health, the economy, and communities at large.

Scientists and researchers raced against the clock to produce effective vaccination frameworks as governments faced numerous challenges posed by this highly contagious and often devastating respiratory virus. These vaccines, lauded as scientific achievements, carry the prospect of minimizing the virus's catastrophic impact and restoring normalcy in our lives. However, the issue was not just about producing the vaccines but also about generating confidence in the public about the vaccine itself. Vaccines play a major role in fighting against the COVID-19 virus, but majority of the people are reluctant owing to many known and unknown reasons. This is vaccine hesitancy, a major obstacle to public health and herd immunity. This thesis will examine the public sentiment towards vaccines through social media analysis and throw light on the major concerns that lead to vaccine hesitancy among the public.

1.1 Relation between social media and COVID-19 Vaccine

Social media is a useful conduit for individuals to acquire information, which helps increase public knowledge and awareness of COVID-19 vaccines (Zhang, 2023) and it is important to note that Social media to a large extent helps form public opinion/sentiment regarding a sensitive topic as vaccination. The effectiveness, safety, and side effects of COVID-19 vaccines are still unknown, so it is critical that individuals have access to the most up-to-date vaccination information. Social media can assist in spreading information swiftly and effectively by removing obstacles in time and space. During the COVID-19 pandemic, an increasing number of people are receiving vaccine advice from social media platforms. People actively resort to social media for news regarding the safety and risks of COVID-19 vaccines given the ongoing concerns about their safety and risks.

However, after the first set of vaccines was rolled out, people were skeptical about how these vaccines might be harmful to their bodies. There were also many conspiracy theories circulating among the general public regarding vaccines, which made people more hesitant to vaccines. Even though vaccine hesitancy has been talked about a lot in recent times, it’s not new. Every time a new vaccine is introduced for the eradication of a certain disease people tend to get anxious about how this product might become harmful for them in the long run. According to the WHO, vaccine hesitancy was considered one of the top 10 threats to global health in 2019 (Cascini, 2022). This information was also spread through numerous social media posts by vaccine-hesitant people or anti-vaxxers are commonly called. On social media platforms such as Reddit, they create their own community, spreading false accusations about vaccines, which negatively influences people.

During a health crisis, social media serves as a platform for exchanging information and opinions and has thus become a critical data source that provides insights into the public's thoughts and attitudes (Kwon and Park, 2023). Unlike traditional approaches such as surveys, which have constraints such as delayed replies, analysing social media conversation has been shown to be an effective way to acquire thorough and timely knowledge of public attitudes based on massive amounts of content. Also, social desirability and recall issues, both of which can limit the validity and reliability of surveys.

Researchers are keen to maximise the number of immunised people with the help of social media posts. Social media data is one of the most important sources of information for researchers seeking to understand public opinion on a variety of topics. For a long time, scholars have relied on Twitter and Reddit, where people freely share their views and these posts can be collected by the researchers as they are freely available for analysis. People share their thoughts on a number of current challenges and triumphs, often with contrasting feelings. Researchers can evaluate people's attitudes from their posts and extract valuable information from them using Sentiment and Semantic Analysis, Topics Modelling and Classification models, proving them to be useful analytical tools. The overall outcome is determined by public opinion: positive, negative, or neutral. The COVID-19 epidemic was one such topic that drew a lot of interest (Kwon and Park, 2023).

1.2 Research aim and objectives

In this research, the main aim is to study public concern over the side effects of COVID-19 vaccines, especially the concern related to myocarditis/pericarditis as a side effect of the COVID vaccine and formulate methods to create vaccine confidence among the public. The research focuses on answering 2 questions based on COVID-19 vaccine hesitancy. The questions are –

RQ1) Impact of Myocarditis compared to other side effects on COVID vaccine hesitancy.

H1) Myocarditis has more impact on public compared to other COVID vaccine side-effects increasing vaccine hesitancy

RQ2) Evaluating public concern over contracting Myocarditis after getting vaccinated by comparing various reports to the results obtained in the social media analysis.

H2) There is a significant correlation between the official statistics on people who have contracted Myocarditis due to the vaccine and the public sentiment on social media platforms regarding the same.

The aims and objectives that will be achieved in this research are as follows:

a) Build a dataset discussing COVID-19 Vaccine hesitancy on Reddit.
b) Compare the impacts of Myocarditis and other vaccination side effects on COVID-19 vaccine hesitancy by looking at the sentiment trends and issues in the Reddit threads.
c) Determine the prevalence of Myocarditis in the dataset by counting the number of entries that list it as a side effect. Achieving this objective will answer the 1st research question RQ1.

d) Examine the relationship between the sentiment nature of Reddit opinions and how it affects the number of people who are hesitant to get vaccinations.

e) Analyse the results and compare them with media reports to gain insights on Myocarditis as a side effect and prove the hypothesis. This objective will answer the question RQ2.

1.3 Dissertation outline

In this dissertation, a critical review of the literature of previous research papers is done explaining issues like vaccine hesitancy, side effects due to vaccines, and the spread of misinformation. Although these papers addressed most of these issues they didn’t focus on any specific side effects like Myocarditis/Pericarditis which are long-term side effects of the COVID-19 vaccine. My research is about evaluating the impact of myocarditis compared to several other side effects caused by the vaccine and comparing the concern of the public from Reddit posts to statistics from official government websites regarding myocarditis. The methodology used for this Topic Modelling, Machine Learning models for labelling the posts retrieved from Reddit using the PRAW library and Keyword Analysis using Word Cloud and bar charts.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we will learn about vaccine hesitancy, and myocarditis/pericarditis as a side effect of covid vaccine and will critically review previous works related to vaccine hesitancy and social media analysis.

2.1 Vaccine Hesitancy

The rising usage of internet health information sources by the public has the potential to influence vaccination uptake (Boucher, 2023). While the internet environment has the potential to increase vaccine knowledge and attitudes about immunization by allowing individuals to share their knowledge and experiences, it also has the ability to spread and amplify misinformation about vaccination.

For instance, in one of the study the author speaks of the Human Papillomavirus (HPV) vaccine which was also refused by the public initially (Dunn, 2015). The vaccine was first approved for use in the United States in 2006 with the goal of lowering the incidence of HPV, which is responsible for the majority of cervical malignancies, as well as genital warts and some oral, anal, and penile cancers. In Australia, HPV vaccination has resulted in a significant reduction in rates of high-grade cervical abnormalities and early signs of herd immunity. However, the uptake of HPV vaccines varies greatly between and within countries. In several countries, the introduction of HPV vaccination was slowed by controversy, with some parents attributing illness or death in their children to the vaccine despite evidence confirming the vaccine's good safety record.

Similarly, in case of the COVID vaccines the hesitancy had to do mainly with the side effects that arise afterwards. There are numerous side effects that are caused by COVID vaccine like Nausea, Fever, muscle and joint pain, etc. But, these problems don’t last longer than a couple of days and thus, won’t cause any major issues to the human body. People are worried of the side effects which might cause them long term health issues such as blood clotting. Even though these are rare conditions that can happen to a patient, the lasting effect on a human body induces fear in peoples’ minds thus hesitating to get immunised. Myocarditis/Pericarditis is one such major side effect caused due to the COVID vaccine.

2.2 Myocarditis/Pericarditis

Myocarditis is an inflammatory disorder of the heart muscle (myocardium) (Voleti, Reddy and Ssentongo, 2022) (Behers, 2022). It can cause symptoms including chest pain, shortness of breath, a rapid or irregular pulse, exhaustion, and, in severe situations, heart failure. While myocarditis can be caused by a variety of factors, including viral infections and autoimmune reactions, myocarditis has been reported as a potential side effect of COVID-19 vaccination. It is crucial to mention that myocarditis after COVID-19 vaccination is extremely infrequent. Cases have been recorded mostly in young boys under the age of 30, and the total incidence remains modest when compared to the millions of doses provided.

The Centers for Disease Control and Prevention (CDC) and other health authorities have actively monitored and investigated these cases in order to assess the risks and benefits of COVID-19 vaccination (CDC, 2020). Most occurrences of myocarditis caused by immunization are minor and resolve with appropriate medical treatment. Even when this unusual adverse effect is considered, the benefits of COVID-19 vaccine in preventing severe illness, hospitalization, and death due to the virus much outweigh the risks. Individuals, particularly those within the recommended age categories for vaccination, should be aware of the possibility of myocarditis as a rare side effect, according to health officials. They stress the necessity of obtaining medical assistance as soon as possible if any worrying symptoms, such as chest pain or palpitations, emerge after receiving a COVID-19 vaccine. Medical personnel are prepared to analyze and handle such cases so that individuals who may suffer this side effect have the best potential outcomes.

2.3 Reddit for social media analysis

Even though there are a lot of social media platforms where people share their views publicly, analysing social media data from Reddit provides distinct research benefits. Also, the main reason for choosing Reddit over a platform like Facebook or Instagram is due to its lack of genuine and relevant data for research purposes (Kwon and Park, 2023). The platform is more “Fun-oriented” and thus data from these platforms can’t be utilised in research. Reddit has numerous subreddits that allow deep and frank debates on a variety of issues, making it useful for gaining a better grasp of specialised interests and community dynamics. The platform's threaded comment system encourages in-depth discussions, which is great for gathering nuanced perspectives. Anonymity on Reddit allows users to share personal experiences and viewpoints that they might not share on more prominent sites. Structured data with upvote/downvote numbers makes tracking content popularity easier. Engaged user communities in a variety of interest areas improve research potential. Because public data is available via Reddit's API, large-scale data collecting for analysis is possible.

2.4 Related Works

Since COVID-19 quickly infected people, there has been an increase in research studies studying the relationship between social media and vaccinations, as well as how users' awareness of the latter was influenced by their usage of the former as a source of information. As a result, noteworthy research conducted from this perspective where researchers have made use of different social media platforms is listed below in a worldwide context.

In a study by Kumar (2021), was regarding the acceptability of the COVID-19 vaccine, emphasizing its vital role in containing the epidemic. It addresses issues like vaccine hesitancy, disinformation, and shifting attitudes. The study investigates the behavior of top users, subject evolution with vaccine events, and mobility among subreddits with varied perceptions in COVID-19 vaccine conversations on Reddit. The results show a complicated link between vaccination incidents, disinformation, and user behavior. Another study that dives deep into vaccine hesitancy, but the purpose of this study is to look into thematic and emotional differences in COVID-19 vaccine side effects discussions on Twitter, Reddit, and YouTube (Kwon and Park, 2023). The authors want to know how different platforms influence user dialogue on vaccination adverse effects. Using content and sentiment analysis, the study identifies common themes and emotional tones in each platform's discussions. In Vishwakarma and Chugh's (2023) study, they use sentiment analysis on Twitter data to assess public perception and consequences of COVID-19 vaccination in India.

Through the analysis of Twitter content, the study intends to assess the sentiment conveyed by Indian society regarding the vaccination. The authors investigate popular opinions, worries, and sentiments towards immunisation using sentiment analysis techniques. The tweets were retrieved from January 2021 to March 2023 and VADER was used for the sentiment Analysis part.

Canaparo (2023) in his paper, describes a novel natural language processing (NLP) methodology for assessing COVID-19 vaccination responses in tweets from several languages and geographical areas. The study addresses the difficulty of determining global public opinion on immunizations. The authors intend to extract insights from multilingual and geolocated tweets about COVID-19 vaccinations using NLP techniques. This technique leads to a better knowledge of worldwide public perceptions, fears, and attitudes toward vaccination, providing vital insights for public health campaigns, communication tactics, and policy decisions in the ongoing pandemic fight. Sussman's (2023) study examines COVID-19 subjects and emotional frames in the context of vaccine hesitancy using social media text and sentiment analysis. The study looks into the themes and emotional tones conveyed in social media discussions about vaccine reluctance. The scientists hope to identify the causes affecting vaccine hesitancy during the epidemic by evaluating the sentiment and content of these exchanges. A study by Liu (2023), examines COVID-19 vaccine sceptics on Chinese social media sites. They investigate the features of various user groups expressing vaccine reluctance, analysing their profiles, interactions, and patterns of involvement. Furthermore, the study dives into the emotion expressed in posts about vaccine hesitancy, with the goal of understanding the underlying mechanisms fueling such sentiment. The authors provide a comprehensive perspective on vaccine hesitancy within the Chinese online community by combining user group analysis and sentiment assessment, shedding light on factors such as vaccine passports influencing public perception and decision-making related to COVID-19 vaccination in this specific social media context.

Catelli (2023) used lexicon-based sentiment analysis to detect opinions and attitudes concerning COVID-19 vaccinations on Twitter in Italy. The researchers examine tweets to better understand how Italian citizens feel about vaccines. Using established sentiment lexicons, the system classifies tweets as good, negative, or neutral depending on the sentiment conveyed in the text. Using this method, the authors can gather insights about public mood and sentiments toward COVID-19 vaccines across the Italian Twitter community. Using sentiment-based topic modelling, Ljaji (2022) hope to find the reasons for COVID-19 vaccine hesitancy in Serbia. The researchers use sentiment analysis to evaluate the emotional tone of vaccination talks, and topic modelling to find the common themes linked with vaccine hesitancy in Serbia. The process entails evaluating text data from multiple sources to extract sentiments and identify significant subjects affecting vaccine hesitancy, providing useful insights into the underlying mechanisms influencing public attitudes and concerns about COVID-19 immunisation in Serbia.

Melton's (2021) research focuses on public sentiment analysis and topic modeling for COVID-19 vaccinations on the social media platform Reddit. The study aims to discover Reddit users' attitudes towards COVID-19 vaccinations and uses topic modelling to detect common themes. The authors stress the importance of these findings, underlining the necessity to take action to enhance vaccine confidence based on insights gained from evaluating feelings and subjects in online discussions, with the goal of addressing vaccine hesitancy and improving public faith in immunization efforts. Here as well LDA has been used to detect relevant topics and for Sentiment Analysis, Textblob is employed.

NLP techniques can be used in analysing any data which contains text and thus, Tunca (2023) used this and Leximancer to analyze the content and sentiment of New York Times stories about the 2019-nCoV (coronavirus). They used VADER in this research for Sentiment Analysis. The study looks at the substance of these stories to identify the issues and themes covered, as well as sentiment analysis to determine the emotional tone expressed in the coverage. The combination of NLP and Leximancer enables a thorough analysis of textual data from The New York Times articles, shedding light on the information landscape and public sentiment during the early stages of the COVID-19 pandemic, which can contribute to a better understanding of media coverage and public reactions to the outbreak. Another study by Özsezer and Mermer's (2022) uses sentiment analysis to evaluate talks about COVID-19 immunization on Twitter in Turkey. The researchers examine the mood expressed in tweets about COVID-19 immunization in Turkey to learn about public opinions, concerns, and overall sentiment toward vaccination initiatives. They employed Machine Learning models for their study namely XGBoost which proves to be the best model. This study provides valuable information on the perception of COVID-19 vaccines within the Turkish online community by examining sentiment trends in these Twitter conversations, which can be useful for designing targeted communication strategies, addressing vaccine hesitancy, and promoting vaccine confidence in the country.

Yan (2021) used Reddit comments to compare public reaction about COVID-19 vaccines in several Canadian cities. In this study, the researchers retrieve data from subreddits which are region specific allowing them to understand the sentiments of public in city-level. LDA was used to determine the topics discussed and Machine Learning models were used to assign polarity for each comment. This analysis gives insights into geographical variances in public views, concerns, and overall sentiment concerning immunization initiatives in Canada by assessing sentiment variations. Verma (2022) study examines COVID-19 vaccination reluctance from an Indian viewpoint using statistical analysis of tweets. The project used opinion mining tools to examine the mood and viewpoints stated in these tweets, with the goal of gaining insights regarding vaccine reluctance in India. As for the methodology, LDA was used to determine the topics discussed and VADER was employed for Sentiment Analysis.

In all the research above, vaccine hesitancy was their main focus with the unified aim of educating the public about the benefits of the COVID vaccines and helping the government enhance vaccine confidence. They even point out the circulation of fake news potentially becoming a global threat. Neither of these papers give insights on any specific long term side effects like blood clots, menstrual cycle issues, etc. Public's reluctance towards vaccines is mainly because of these long term side effects which might affect their health negatively, with the worst case scenario being decreased life expectancy. It's the same with a lot of other social media research conducted. Therefore, the aim of this research is to gain insights on the general public opinion on vaccines and their side effects (especially long term side effects). In the final analysis, real world statistics on Myocarditis, a major long term side effect, will be compared to the social media analysis outcome obtained so that the public can be educated about various misconceptions about the vaccine side effects. The statistics will be obtained from government websites like MHRA (which operates the

Yellow Card Scheme for COVID-19 vaccine-related issues) regarding myocarditis caused due to vaccines.

2.5 Summary

In this chapter, we learnt discussed in detail about vaccine hesitancy and the reasons behind this. Also, a brief explanation on what Myocarditis is and how people are reacting towards this side effect has been explained considering this is my main topic. How social media analysis has helped in gaining useful insight to build vaccine confidence among people and the use of Reddit for social media analysis. There was also critical review on the literature of previous works related to social media analysis and vaccine hesitancy. In this thesis, the main focus will be on solving the Research question of how big of an impact does myocarditis have over other side effects and public’s concern on the same which is the research gap mentioned in this chapter.

CHAPTER 3: METHODOLOGY

For answering the Research Questions and achieving the objectives mentioned before, the following steps will be implemented as follows:

1. Data Collection:

We will collect the data related to the COVID vaccine using the Reddit API, then using the PRAW (Python Reddit API Wrapper) library in Python to access the posts, and then collect only those which are relevant for answering the research question based on the keywords related to the same. For example, posts that are related to vaccine hesitancy from subreddits like r/CovidVaccine, r/Coronavirus, etc. can be extracted with keywords like ‘’myocarditis”, “vaccine hesitancy”, “side effects”, etc.

2. Data Pre-processing:

Pre-processing of the data is an important aspect of removing the irregularities in the dataset and achieve maximum accuracy in the model. This is done by removing the stopwords, URLs, and special characters from the dataset. Case uniformity is an important part where all the words should be converted into lowercase. Lemmatization, Tokenization, Parts of Speech Tagging and Named Entity Recognitions are implemented in this section. The 1st objective of the dissertation is fulfilled with data collection and preprocessing steps.

3. Feature Extraction:

Techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and CountVectorizer are used to transform the preprocessed text into numerical features to ensure they are suitable for Machine Learning algorithms. These methods are also used for applying Topic Modelling.

4. Topic Modelling:

Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) are implemented in this section to recognize the topics that are discussed in the posts other than COVID-19 vaccines. The model that produces better results will be used to remove the posts that are not relevant to vaccine hesitancy.

5. Labelling Vaccine Hesitancy:

First, a few 100s of posts will be labelled manually into 2 categories - “Vaccine Hesitant” and “Vaccine Non-hesitant”. These posts will then be trained using Machine Learning algorithms which will be then used to label the rest of the Reddit posts. The ML model which has the highest accuracy will be the one used for classifying the posts. The evaluation will be analyzed using different metrics such as F1 score, precision, and recall.

6. Sentiment Analysis:

Perform sentiment analysis on the Reddit posts using a sentiment analysis library, VADER(Valence Aware Dictionary for Sentiment Reasoning), to capture the sentiment expressed towards COVID vaccines. Explore the relationship between sentiment and vaccine hesitancy to understand the sentiment patterns associated with hesitancy. This step aligns with objective 4 of the Aims and Objectives..

7. Identifying vaccine hesitancy reasons:

The frequency of the words will be calculated to understand the reasons the public is mainly concerned about and then sort them based on their frequencies to get the topics discussed. We can also generate visualisations to represent this data. The NLTK library will be used to get the term frequencies. Objectives 2 and 3 of the dissertation is fulfilled with this step as there will be visualisations used to understand myocarditis as a side effect of the vaccine and its impact.

8. Influence of Myocarditis on Vaccine Hesitancy with Visualisations:

Filter posts specifically discussing myocarditis as a side effect of COVID vaccines. Analyze the influence of myocarditis-related sentiments and concerns on overall vaccine hesitancy sentiments. Visualisations (e.g., bar charts, word clouds) will be created to present the findings, including reasons for vaccine hesitancy, sentiment patterns, and the influence of myocarditis. The results obtained from the visualisations will provide insights on Myocarditis as a side effect of the vaccine. With this step the final objectives is accomplished and with it the research questions are answered.

9. Ethical Approval

Reddit is a social media platform where people share their views freely without stressing about privacy. The platform allows users to view anything that is posted by anybody in the world. The same goes for researchers as they are free to parse the posts of public without any privacy concern. And to make things ethically clean all these posts are gone through preprocessing steps so as to eliminate any personal contacts or identity of the person. The posts are parsed as a whole rather than individually. Also, I am not allowed to write any posts in this report unless and until they are paraphrased so that they are not identified due to their posts. Thus, according to the Ethics Department, I was free to conduct my research without the need for ethical approval.

CHAPTER 4: RESULTS

This module focuses on the outcomes of each step that were used to obtain the desired results regarding Myocarditis as a side effect of COVID vaccines and its impact in terms of vaccine hesitancy. The reason why certain methodology was used for this analysis over the others in previous research and the output obtained is explained in detail. This chapter will lay the foundation for the analysis that will be covered in the next chapter.

Data collection and preprocessing is the most important part of Data Analysis. Irregularities in the data or irrelevant data often produces misleading outcomes leading to conclusions which are far off from the expected results. In this research, it was made sure that the posts from the Reddit social media were genuine and can be trusted for the analysis purpose. People post their views and opinions in the Reddit platform through subreddits which are similar to channels on a Television where each channel is dedicated for a particular subject. There are thousands of subreddits depending on your topic of interest but not all of them will have genuine information which is concerning from an analysis Point of View. Thus, choosing the subreddits on COVID-19 related topics which have valuable yet legitimate information or rather posts from users who share genuine information on such platforms.

For the collection of data regarding COVID vaccine hesitancy, I shortlisted the subreddits which were popular among the users. For that, I looked at the number of subscribers that each subreddit had garnered over a certain period of time. The threshold of number of subscribers was at least 1000 users which reveals the genuinity of the subreddit. Also, all the subreddits from where the posts were retrieved promoted the users to share their experiences regarding their vaccine intake and also guide others who are hesitant towards the vaccines. The list of all the subreddits are as follows: Antivax, changemyview, Coronavirus, CoronavirusDownunder, CoronavirusUK, COVID19, COVID19_support, COVID19Europe, COVID19positive, CovidVaccinated, VACCINES, Vaccine, CovidVaccine and ScienceBasedParenting (Note: All these names have “r/” in the front indicating they are subreddits whereas “u/” means they are users of the Reddit platform).

For extraction of the posts from Reddit, Python Reddit API Wrapper (PRAW), a package used in Python programming language, has been used. The posts were retrieved from each of the above-mentioned subreddits using specific keywords like myocarditis, side-effects, vaccines, coronavirus, covid-19, antivaxxers, conspiracy, anti-vaccination, hesitant, etc. The features that were used in this were mainly Title, Selftext (Content column in my case) and Score of each post. After retrieval, Pandas library is used to create a dataframe with the columns as mentioned and then saved as a Comma Separated Values (CSV) file. These CSV files of different subreddits were then concatenated to form a single CSV file with 6486 records of social media content. In some records, the Content column was Null thus replacing those values with their corresponding Title and then dropping the Title column considering the information relevant to us will be in the Content column of the dataframe. While the chance of having repetition of posts in different subreddits was high, they were dealt with by using the ‘drop_duplicates’ method.

The next step is Data Preprocessing which is the most time-consuming part of all the steps. Firstly, every post that is retrieved from Reddit starts with the letter ‘b’ and is enclosed in double inverted commas. These are removed using the ‘strip’ function along with the use of a python package called ‘redditcleaner’ which is specifically designed for Reddit text analysis to remove irregularities such as URLs, newlines, strike through etc. For the text analysis to be accurate, it is important to convert capital letters to lowercase. Removing punctuations, contractions, emoticons, numbers and stopwords are also an important step for the analysis. Stopwords like ‘not’ and ‘side’ were removed from the list of stopwords considering these words will be important while creating ‘Ngrams’ like ‘side effects’ and ‘not hesitant’. Tokenization (a process of parsing larger amounts of text to smaller parts) and Lemmatization (converts the tokens to their meaningful root forms such as Giving => Give) are also essential for the same. These processes are implemented using the ‘SpaCy’ package in python. SpaCy is an alternative to the rather conventional package Natural Language Toolkit (NLTK). The main advantage of using SpaCy over NLTK is the processing speed and the accuracy it offers for NLP tasks. This library was also used to identify Named Entity Recognition (NER) and for Parts of Speech tagging of the tokens in the dataframe after Tokenization.

Subreddits like “changemyview” and “ScienceBasedParenting” are not COVID-19 specific information forums. Their platform is used to share anything which is related to science and mythbusters. Thus, after data collection, it was understood that there were some records which were irrelevant from the research Point of View. I implemented Topic Modelling techniques like Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) to separate posts into 5 different topics so that the posts which were not relevant for my research can be dropped from the dataframe. From the outputs, it was evident that the division of words based on the tokens into separate topics was much more accurate using NMF when compared to the LDA model. For the implementation of these models, the features of the dataframe needs to be converted to numerical form and for that Term Frequency - Inverse Document Frequency (TF-IDF) Vectorizer and Countvectorizer are used for NMF and LDA respectively. After implementation of the following models, about 552 records were found to be not related to COVID vaccines and therefore they were excluded from the new dataset created after cleaning with the final count of 5931 records. LDA was used from the Scikit learn library instead of Gensim even though latter is better in terms of scaling large amounts of text corpuses. But NMF, which is better than LDA when it comes to producing meaningful topics can’t be imported using Gensim library and thus, it is easier to operate both using sklearn.

Next step is labelling the posts in the cleaned dataset where I labelled the first 198 records into vaccine hesitant (H- 64) and vaccine non-hesitant (N- 134) posts. These posts were used to train a Machine Learning model and then use it to automate the process of labelling the rest of the dataset. For this step, 3 ML algorithms were employed namely Random Forest Classifier, Multinomial Naive Bayes and Logistic Regression. After training and testing the model, the Multinomial Naive Bayes turned out to be better in terms of accuracy of predicting the labels and thus was preferred over the other 2 for labelling the dataset. We were able to get a nearly balanced dataset with 2533 out of 5931 posts being vaccine hesitant and the rest, vaccine non-hesitant.

Metrics Models	Accuracy	Precision	Recall	F1_score
Logistic Regression	0.65	0.74	0.65	0.67
Multinomial Naïve Bayes	0.75	0.76	0.75	0.75
Random Forest Classifier	0.75	0.73	0.75	0.73

Table 1: ML algorithm performance comparison

After labelling, Valence Aware Dictionary and sEntiment Reasoner (VADER) was employed for Sentiment Analysis. We get a column with each document having values negative, positive, neutral and compound (determines the degree of sentiment ranging from -1 to 1). The compound value in the column is used to indicate the overall sentiment of a post where a column named ‘Sentiment’ is created in the dataframe. The value is assigned as 1 for compound values greater than or equal to 0 and for the values less than 0 is assigned as 0. A heatmap is created to get insights on how many records there are for each combination of 2 binary columns namely “Vac Hes/Non” (Vaccine hesitancy ‘H/N’) and Sentiment (1/0). The sentiment of vaccine non-hesitant users were more than hesitant users, but 1778 of them had positive sentiment towards vaccines compared to 1620. The negative sentiment among the vaccine hesitant group is less when compared to other groups but still is more than 1200 out of 5931 user opinions. Here’s the heatmap showing the distribution for each category of vaccine hesitancy and sentiment scores:

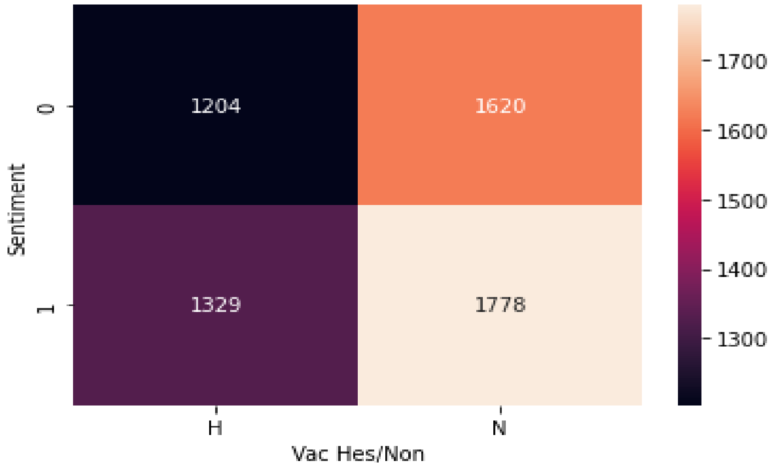


Figure 1. Sentiment vs Vaccine hesitant/Non-hesitant heatmap

4.1 Answering Research Question 1

For answering the first research question RQ1, evaluating the impact of myocarditis to different side effects caused by the COVID vaccine, keyword analysis is implemented. Python packages namely “collections” and “wordcloud” are used and for implementing these, textual data needs to be numerical. For this purpose, TF-IDF vectorizer is used to transform the new dataframe. There are some phrases which are used quite extensively in a lot of posts such as ‘side effect’, ‘not hesitant’, ‘covid vaccine', etc. Thus, to make sure that they are also included in the analysis, ngrams are used with a range of 1 to 3 words i.e. it will include unigrams, bigrams and trigrams.

The analysis is done separately for vaccine hesitancy labels "H" and "N" so that we can gain more information on the concern of people for different groups. Both the Bar graphs and Word Clouds are plotted with respect to the keywords and their tf-idf scores. While plotting these charts, the overlapping of the same words in the form of both unigrams as well as bigrams made it difficult to get an accurate result. Thus, a function was defined to remove those overlapping of the unigarms. But this function not only stopped the overlapping but removed the unigrams altogether. Thus, a list of all the important keywords were introduced to the function so that there will be a balance in both unigrams as well as bigrams.

In the bar graph, it is visible that the terms 'antivaccine', 'hesitancy', 'conspiracy', 'pain', 'myocarditis', etc. are the keywords that have higher tf-idf scores. The same is true for the word cloud as well where there are also bigrams such as 'covid vaccine' and 'side effect' that have high tf-idf scores are spoken about by the Reddit users. Talking about the side effects, other than myocarditis, most of the side effects like 'blood clot', 'nausea', 'sore arm', 'body ache', etc. have low scores which is why their size is smaller than other keywords. In the vaccine hesitant charts, the term 'antivaccine' has been at the top whereas in the non-hesitant chart even though it is in the top 10 it is less spoken about by the users. There are also mentions of vaccine brands like 'pfizer', 'astrazeneca', 'moderna', 'jj vaccine' and vaccine terms like 'mrna' and 'booster'. Among these terms, 'pfizer', 'astrazeneca' and 'mrna' have higher scores. The size of the keywords in the word cloud is directly proportional to their respective tf-idf scores. The scores of few of the words are shown in the bar plots and word clouds below in figure 2, 3, 4 and 5:

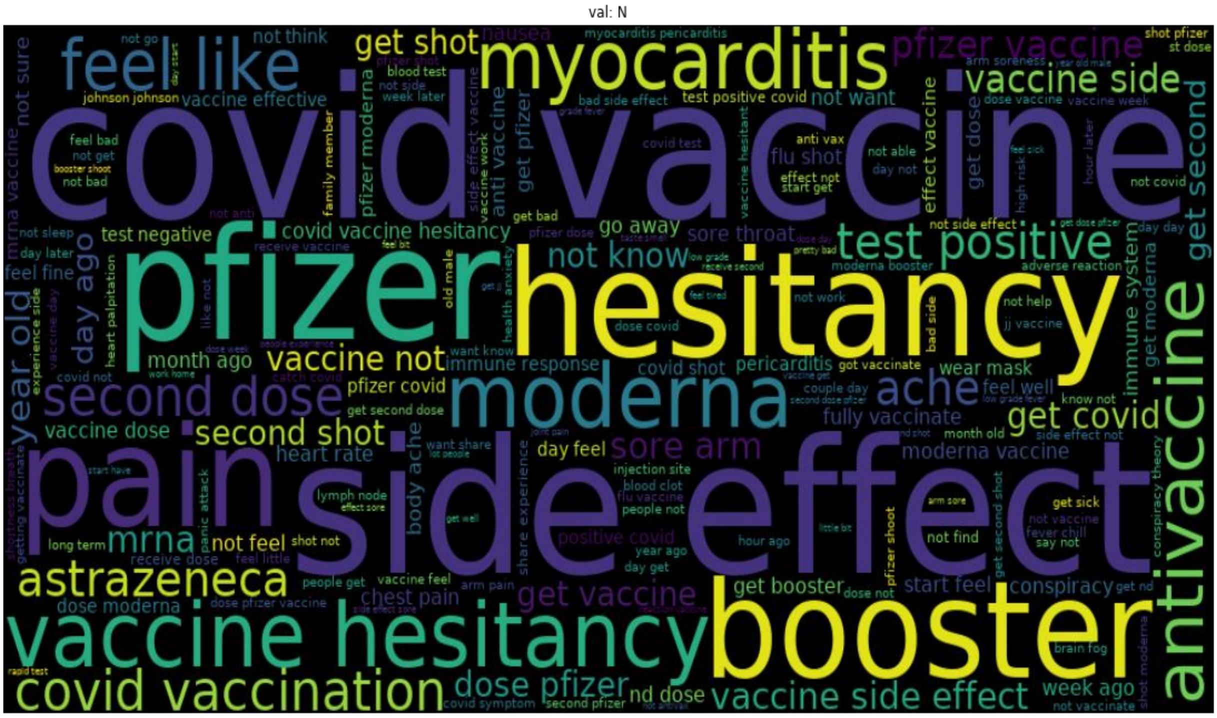


Figure 2: WordCloud for vaccine non-hesitant category

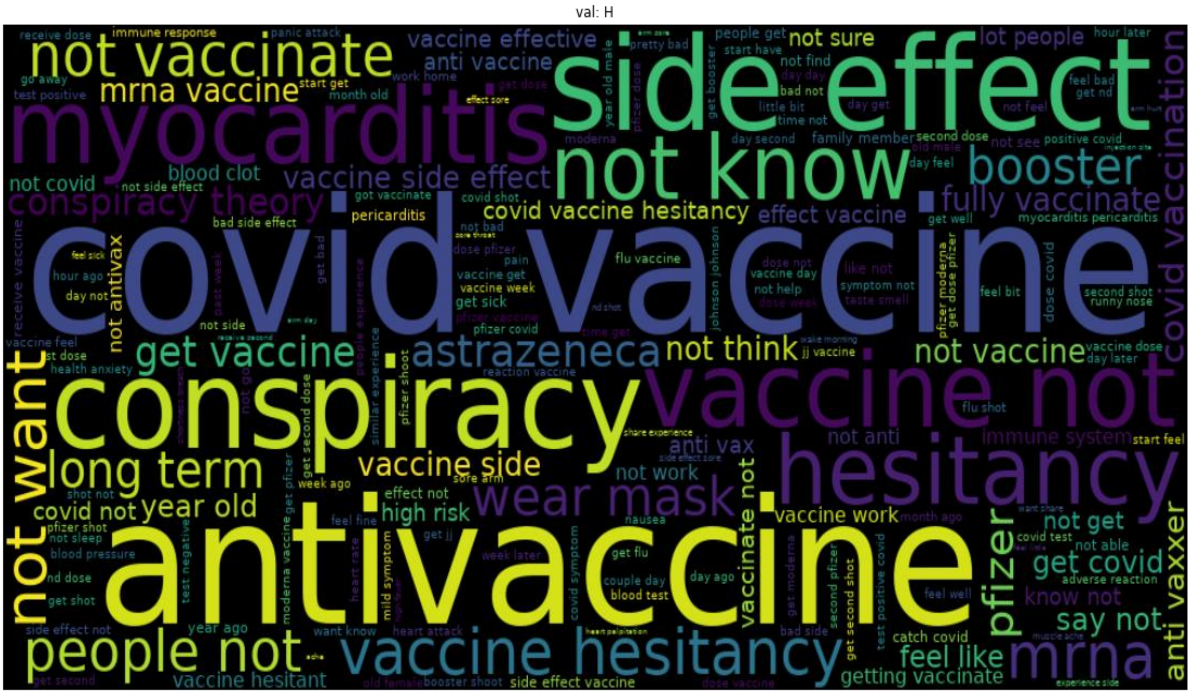


Figure 3: WordCloud for vaccine hesitant category

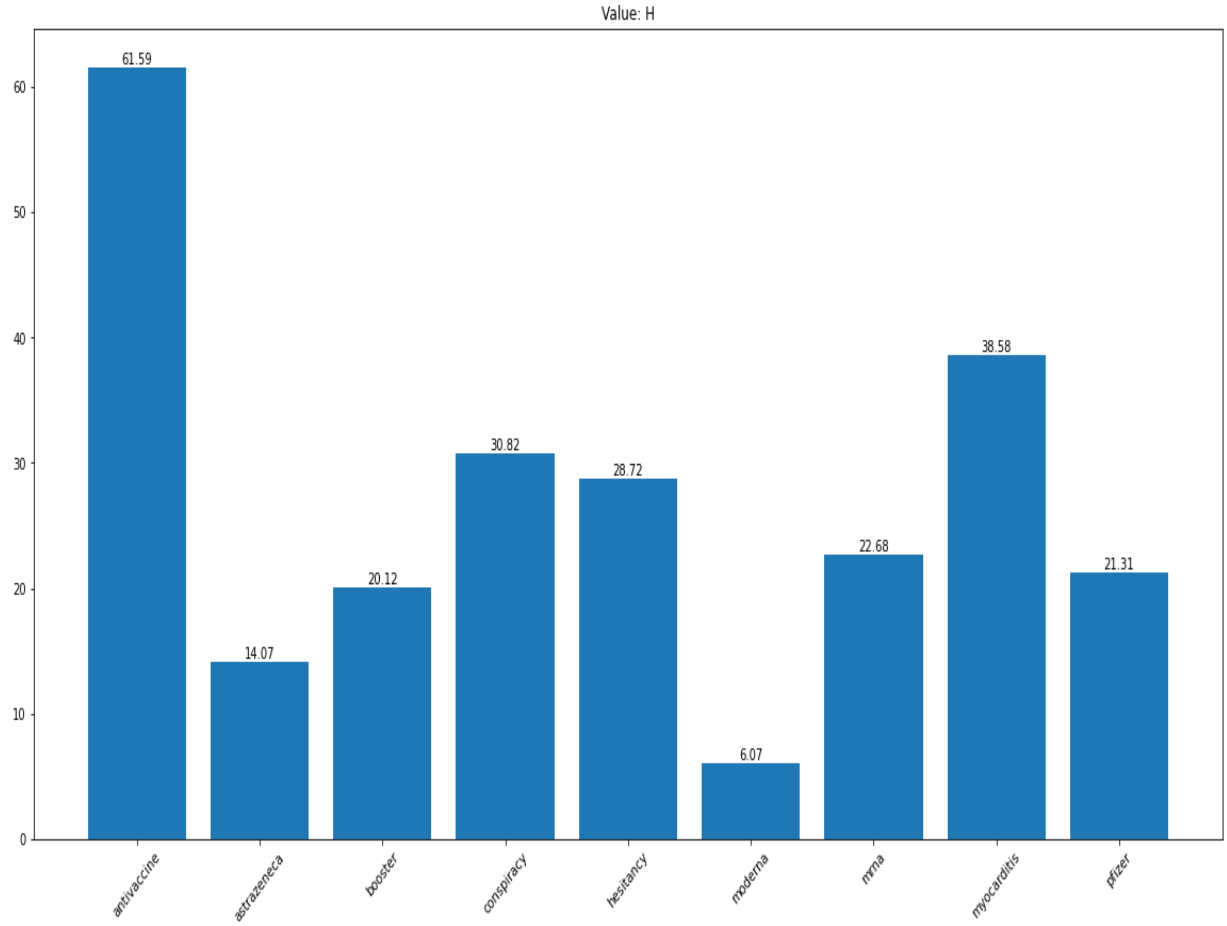


Figure 4. Keyword analysis using Bar plot for vaccine hesitant category

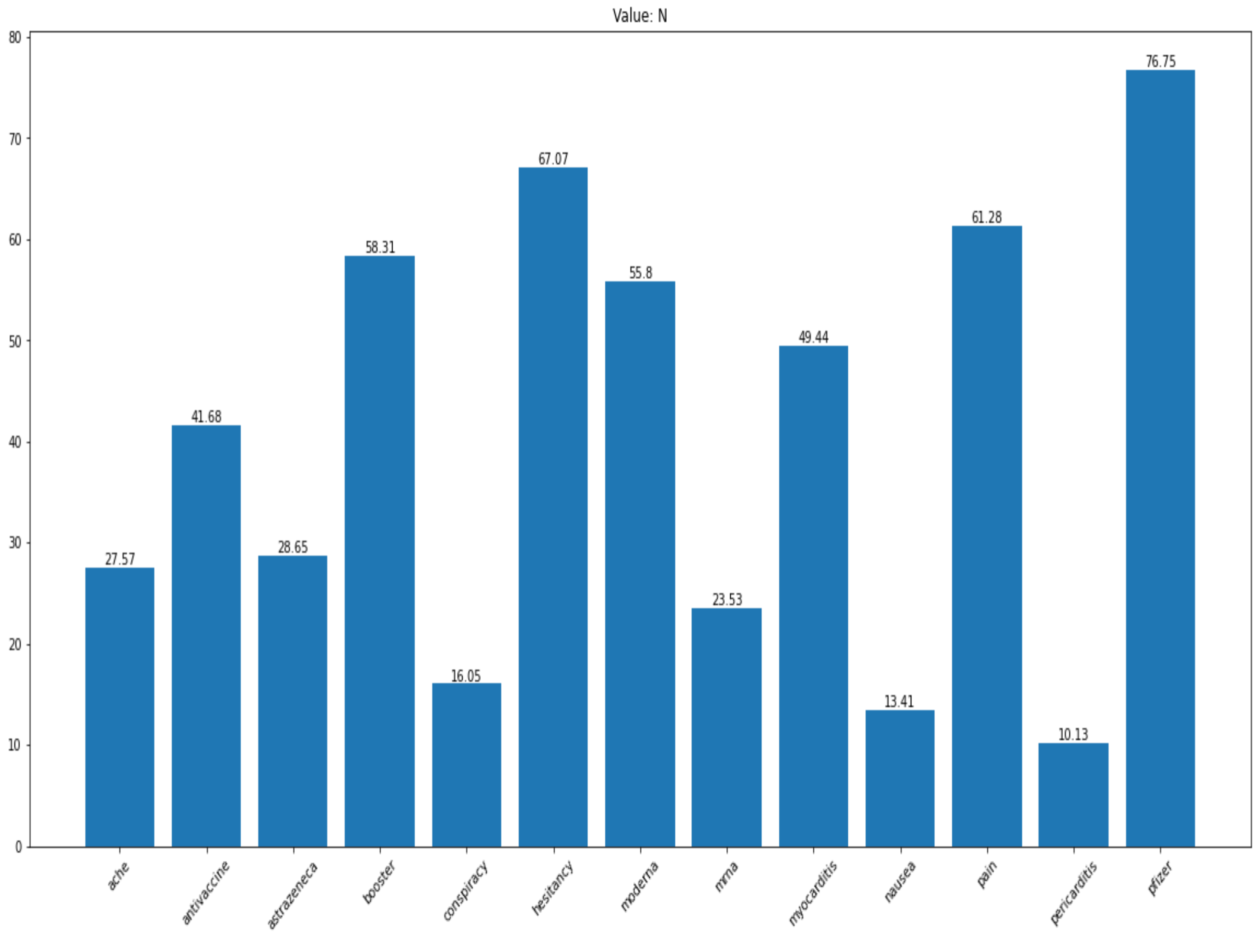


Figure 5. Keyword analysis using Bar plot for vaccine non-hesitant category

4.2 Answering Research Question 2

To answer the second question, only posts which consist of the keyword ‘myocarditis’ or ‘pericarditis’ are extracted from the main dataframe. The resulting dataframe has 353 records with 4 columns namely ‘Content’, ‘Tokens’, ‘Vac Hes/Non’ and ‘Sentiment’. Similar to how a heatmap was created after the VADER Sentiment Analysis, a heatmap for this dataframe is created. From the heatmap, as the numbers from the figure below suggest, people who were non-hesitant were the ones speaking about this specific side effect which implies either they have contracted the side effect after they got vaccinated and are sharing their experience on the same or that they are concerned whether they will suffer from myocarditis/pericarditis after they took the jab. Either way, myocarditis has been spoken of by the ones who have vaccinated or are keen to take the vaccination. Also, the number of vaccine non-hesitant Reddit users are 239 out of 353 total users which is substantially more than the vaccine hesitant category and 139 of the 239 users’ opinions are negative (0). This result will be compared to statistics obtained from different sources in the next chapter. The visual representation for the following is given below:

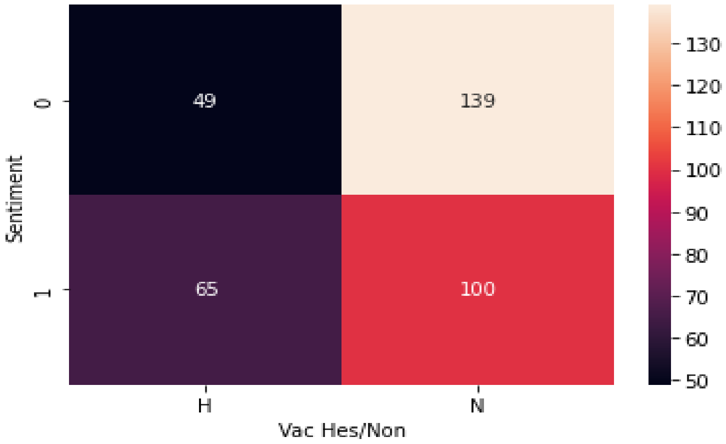


Figure 6. Heatmap showing vaccine hesitancy catgories vs Sentiment score for data containing keyword “Myocarditis or Pericarditis”

4.3 Summary

The main objective of this chapter is to give a detailed explanation on the methodology that I used and the outputs that I obtained in the form of visualisations using keyword analysis. From the Results it is evident that the majority of the the users mentioning the keyword ‘myocarditis’ are vaccine non-hesitant. Also, talking about the impact of myocarditis when compared to other side effects is higher. These outputs will be used for further analysis in the next chapter Discussions.

CHAPTER 5: DISCUSSIONS

In the previous chapter, we were able to learn about how the methodology helped me in getting the results in the form of charts and figures and how they will work as a foundation for the analysis that will be discussed in this chapter. We will also talk about the limitations of the algorithm used which might have affected the overall analysis.

The first research question RQ1 was answered in the previous chapter considering among the side effects, Myocarditis was the only one which had garnered a bigger tf-idf score when compared to other COVID vaccine side effects. There were mentions of side effects such as blood clots, sore arm, flu, body pain, disruption in the menstrual cycle of women (only the keyword "menstrual" is mentioned during keyword analysis), etc. Even though blood clots and menstrual changes are major side effects, from the analysis it was interesting to see people being more concerned towards myocarditis which is one of the least probable side effects to occur due to the vaccine. Thus, the hypothesis H1 "Myocarditis has more impact on public compared to other COVID vaccine side-effects increasing vaccine hesitancy" is accepted. Figures 2 and 3 serve as the evidence for the same.

Myocarditis/pericarditis has been a bit of a concern from the public's point of view and it is evident from what I was able to deduce through the social media analysis. The keyword analysis was clearly able to depict the use of the word 'Myocarditis' in the posts retrieved from Reddit. Also, it's not just Myocarditis, there have also been people mentioning about 'chest pain' or 'heart rate' in these posts (Figure 2 and 3). Both chest pain and increase in heart rate are symptoms of Myocarditis and sometimes due to the complexity of the term itself people mention only the symptoms of such ailments. Also, talking about the term, Pericarditis, which is another term mentioning the inflammation of the outer lining of the heart, is also one of the side effects mentioned by the users.

Reports from UK government's Medicines and Healthcare products Regulatory Agency (MHRA) (coronavirus-yellowcard.mhra.gov.uk, 2022) states that even though initially they had higher number of reports regarding Myocarditis/pericarditis as a vaccine side effect, these reports were very rare and those who did contract this condition recovered rather quickly with due to proper treatment and rest. Also, another pattern that was found regarding the vaccines were that the mRNA vaccines caused these side effects more than other 2 types namely Vector and Protein Subunit vaccines. Both Pfizer and Moderna are mRNA vaccines and according to the MHRA reports there is a high probability of contracting myocarditis because of these vaccines compared to Astrazeneca (Vector) and Novovax (Protein Subunit). The MHRA operates the Yellow Card System which is a scheme which records any or every suspected safety concern occurring due to vaccines and because of this scheme one can gain all vaccine related information in the government's Yellow Card Website. The information that we received from this website is shown below:

The pattern about mRNA vaccine causing side effects is also visible in the keyword analysis where keywords like 'mrna' , 'moderna' and 'pfizer' were extensively used in the Reddit posts. Also, these keywords are in the top 5 in both Hesitant and Non-hesitant categories of users. There were also a few posts where the user mentions how they contracted long term side effects of pericarditis due to the Pfizer vaccine. Also, according to that user, the symptoms started to show in less than a week after the vaccine jab which is the average time period mentioned by Yellow Card System for 2nd dose of Pfizer/BioNTech vaccine. Also, talking about the impact of myocarditis the highest probability of contracting the illness is in the age group of 18-29 years old and the majority of the people are men. The likelihood of contracting Myocarditis/Pericarditis doesn't end here as the reports suggest that on average for all the vaccines the reports per million doses of vaccines is less than 10 for 1st, 2nd, and booster or 3rd doses. Tables related to Myocarditis from the Yellow Card System are given below :

Age Range (Years)	Pfizer			Moderna			Astrazeneca	
	1st dose	2nd dose	3rd dose	1st dose	2nd dose	3rd dose	1st dose	2nd dose
Under 18	13	8	Not calculated	N.A	N.A	N.A	N.A	N.A
18-29	24	29	17	61	70	20	10	17
30-39	20	25	16	60	51	20	14	12
40-49	20	19	13	48	30	16	14	10
50-59	11	18	8	Not calculated	Not calculated	8	8	8
60-69	5	14	7	Not calculated	N.A	8	7	6
70+	4	5	4	Not calculated	N.A	1	4	5

Table 2: By patient age and dose, up to and including 23 November 2022, reporting rates per million doses for UK ADR reports of suspected myocarditis and pericarditis connected with COVID-19 Vaccines.

Age Range (Years)	Pfizer	Moderna	Astrazeneca
Under 18	83	0	0
18-29	396	127	31
30-39	323	98	49
40-49	150	53	123
50-59	110	24	108
60+	168	23	109
Unknown	161	38	52
Total	**1391**	**363**	**472**

Table 3: By patient age, up to and including 23 November 2022, the number of UK ADR reports including suspected myocarditis, pericarditis, and other relevant terminology received for the COVID-19 Vaccines.

Sex	Number of Reports		
	Pfizer	Moderna	Astrazeneca
Female	546	119	212
Male	799	234	250
Unknown	46	10	10
Total	**1391**	**363**	**472**

Table 4: Number of suspected myocarditis, pericarditis, and other heart-related ADR complaints received in the UK for the COVID-19 vaccines, broken down by patient sex up to and including 23 November 2022.

Centres for Disease Control and Prevention (CDC, 2020), also mentioned about the rarity of contracting this disease due to vaccination. According to them, the age group of 12-17 years males experienced myocarditis within a span of 21 days after their 2nd jab of vaccine. The likelihood of contracting the disease is about 22 to 36 per 100,000 patients. Whereas the probability of suffering this illness ranges from 50 to 60% per 100,000 COVID infected males in this specific age group.

In one of the research papers, the researcher (Voleti, Reddy and Ssentongo, 2022) has said about how myocarditis has been dangerous for people who have been infected with COVID-19 rather than those who have been vaccinated. According to the report, a person is likely to suffer from myocarditis/pericarditis is 7 times higher if they are infected by COVID-19 compared to the people who are COVID vaccinated. It is true that there are people who suffered from this disease after getting vaccinated but still the numbers are quite low and factually speaking the fatality rate due to myocarditis is also less than 20 from the Yellow Card System’s report from December 2020 to November 2022.

Another point to be mentioned is regarding the Data Analysis part where in the heatmap, vaccine non-hesitant people are the majority by quite a margin. The number of vaccine hesitant people who have mentioned myocarditis in their posts are only 114 out of 353 (Figure 6) users out of which the sentiment of 65 people are positive which means they might not be against vaccines but their concerns regarding the same persists. From the tf-idf score of each word one can understand that the concern regarding Myocarditis (highest tf-idf score among side effects) is genuine and there is strong correlation between the social media analysis and the statistics provided by Government organisations like MHRA and CDC. Thus, answering the research question RQ2 and accepting hypothesis H2 based on the Figures 4 and 5 and the statistics from MHRA and CDC respectively.

CHAPTER 6: CONCLUSION

N.1 Summary of the dissertation

In this dissertation, the main motive is to talk about the side effects of the COVID-19 vaccine and vaccine hesitancy by analyzing the data from social media platforms like Reddit. In the Literature Review, a lot of previous research works were analysed and from that, we realized that these researches haven't focussed on a particular side effect which is the main highlight of this thesis. For the research purpose, a methodology was fine-tuned to get an accurate result from the analysis. The Results obtained regarding Myocarditis as a side effect of the COVID-19 vaccine were discussed using NLP methods like Topic Modelling, keyword analysis, and Machine Learning models and the output is then compared to statistics from government websites to evaluate the impact of the side effect. From the outcomes obtained, both the research questions were answered, and the hypotheses were accepted after analysis.

N.2 Research contributions

In this research, the main objective is to promote vaccine confidence among the public. Due to a lot of misinformation spreading across social media, like side effects of vaccines that were not even found or reported officially, and various conspiracies increased vaccine hesitancy. So, it is important to educate people regarding the side effects of the COVID-19 vaccines like how it was proven that Myocarditis, even though a long-term side effect, is the least probable side effect. Also, even if you do contract the illness, timely treatment and rest can cure it in a few days. It's through such research, that people will realize the importance of vaccines.

N.3 Future research and development

The methodology that was used in this thesis was chosen considering the accuracy and performance for data analysis purposes. But, as technology tends to become outdated after a period, the same could be said about the methodology which even though it uses some of the best-performing algorithms such as SpaCy, redditcleaner, NMF, and Scikit Learn might not prove to be the best in the future. On the bright side, this brings forth new opportunities for future research works.

- My research solely focussed on data from Reddit social media. For future works, one can use a hybrid model where data from 2 or 3 social media platforms can be used to get a wide variety of data. Collecting data from different sources will cover more demographics and thus will get a more comprehensive view.

- Vaccine hesitancy although an intriguing topic won't pique anybody's interest if there is nothing new as a few years down the line researchers should look for something new that might have evolved from this topic. Like the evolution of vaccine hesitancy or the impact of the introduction of new vaccines.

- This thesis focussed only on a single language but it can be done for multiple languages from social media which will make it much more interesting.

- Development of new technology and new techniques means that state-of-the-art techniques used at present will soon become old and inefficient for future analysis.

N.4 Personal Reflections

One of the strengths that I have is analyzing the results which not only in academics but it has always been one of my strengths to analyse any given situation and predict an outcome. I have used this strength for predicting results for sporting events by engaging in SWOT analysis. Speaking of weakness, I like everything to be done perfectly which sometimes backfires because while looking for perfection sometimes I lose track of time which gives me unnecessary stress and anxiety. But, during this dissertation, I have tried to accept the results that I have obtained and made myself realise that I can improve on this in the future. So in the future, I will have to make sure that I won't take unnecessary stress by going for perfection and ending up with a result that is worse than expected.

REFERENCES

1. Melton, C.A., White, B.M., Davis, R.L., Bednarczyk, R.A. and Shaban-Nejad, A. (2022). Fine-tuned Sentiment Analysis of COVID-19 Vaccine–Related Social Media Data: Comparative Study. *Journal of Medical Internet Research*, 24(10), p.e40408. doi:https://doi.org/10.2196/40408.
2. Zhang, Q., Zhang, R., Wu, W., Liu, Y. and Zhou, Y. (2023). Impact of social media news on COVID-19 vaccine hesitancy and vaccination behavior. [online] 80, pp.101983–101983. doi:https://doi.org/10.1016/j.tele.2023.101983.
3. Cascini, F., Pantovic, A., Al-Ajlouni, Y.A., Failla, G., Puleo, V., Melnyk, A., Lontano, A. and Ricciardi, W. (2022). Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature. *eClinicalMedicine*, [online] 48, p.101454. doi:https://doi.org/10.1016/j.eclinm.2022.101454.
4. Kwon, S. and Park, A. (2023). Examining thematic and emotional differences across Twitter, Reddit, and YouTube: The case of COVID-19 vaccine side effects. *Computers in Human Behavior*, 144, p.107734. doi:https://doi.org/10.1016/j.chb.2023.107734.
5. Boucher, J.-C., Kim, S.Y., Jessiman-Perreault, G., Edwards, J., Smith, H., Frenette, N., Badami, A. and Scott, L.A. (2023). HPV vaccine narratives on Twitter during the COVID-19 pandemic: a social network, thematic, and sentiment analysis. *BMC Public Health*, [online] 23(1). doi:https://doi.org/10.1186/s12889-023-15615-w.
6. Dunn, A.G., Leask, J., Zhou, X., Mandl, K.D. and Coiera, E. (2015). Associations Between Exposure to and Expression of Negative Opinions About Human Papillomavirus Vaccines on Social Media: An Observational Study. *Journal of Medical Internet Research*, [online] 17(6), p.e144. doi:https://doi.org/10.2196/jmir.4343.
7. Voleti, N., Reddy, S.P. and Ssentongo, P. (2022). Myocarditis in SARS-CoV-2 infection vs. COVID-19 vaccination: A systematic review and meta-analysis. *Frontiers in Cardiovascular Medicine*, 9. doi:https://doi.org/10.3389/fcvm.2022.951314.
8. Behers, B.J., Patrick, G.A., Jones, J.M., Carr, R.A., Behers, B.M., Melchor, J., Rahl, D.E., Guerriero, T.D., Zhang, H., Ozkardes, C., Thomas, N.D. and

Sweeney, M.J. (2022). Myocarditis Following COVID-19 Vaccination: A Systematic Review of Case Reports. *The Yale journal of biology and medicine*, [online] 95(2), pp.237–247. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9235262/.

9. CDC (2020). *COVID-19 Vaccination*. [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/myocarditis.html.
10. Kumar, N., Corpus, I., Hans, M., Harle, N., Yang, N., McDonald, C., Sakai, S.N., Janmohamed, K., Tang, W., Schwartz, J.L., Jones-Jang, S.M., Saha, K., Memon, S.A., Bauch, C.T., De Choudhury, M., Papakyriakopoulos, O., Tucker, J.D., Goyal, A., Tyagi, A. and Khoshnood, K. (2021). *COVID-19 vaccine perceptions: An observational study on Reddit*. [online] doi:https://doi.org/10.1101/2021.04.09.21255229.
11. Canaparo, M., Ronchieri, E. and Scarso, L. (2023). A natural language processing approach for analyzing COVID-19 vaccination response in multi-language and geo-localized tweets. *Healthcare Analytics*, [online] 3, p.100172. doi:https://doi.org/10.1016/j.health.2023.100172.
12. Anjali Vishwakarma and Chugh, M. (2023). COVID-19 vaccination perception and outcome: society sentiment analysis on twitter data in India. *Social Network Analysis and Mining*, 13(1). doi:https://doi.org/10.1007/s13278-023-01088-7.
13. Sussman, K.L., Bouchacourt, L., Bright, L.F., Wilcox, G.B., Mackert, M., Norwood, A.S. and Allport Altillo, B.S. (2023). COVID-19 topics and emotional frames in vaccine hesitation: A social media text and sentiment analysis. *DIGITAL HEALTH*, 9, p.205520762311583. doi:https://doi.org/10.1177/20552076231158308.
14. Liu, J., Lu, S. and Zheng, H. (2023). Analysis of Differences in User Groups and Post Sentiment of COVID-19 Vaccine Hesitators in Chinese Social-Media Platforms. *Healthcare*, 11(9), p.1207. doi:https://doi.org/10.3390/healthcare11091207.
15. Catelli, R., Pelosi, S., Comito, C., Pizzuti, C. and Esposito, M. (2023). Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy. 158, pp.106876–106876. doi:https://doi.org/10.1016/j.compbiomed.2023.106876.

16. Ljajić, A., Prodanović, N., Medvecki, D., Bašaragin, B. and Mitrović, J. (2022). Uncovering the Reasons behind COVID-19 Vaccine Hesitancy in Serbia: Sentiment-Based Topic Modeling . *Journal of Medical Internet Research*. doi:https://doi.org/10.2196/42261.

17. Melton, C.A., Olusanya, O.A., Ammar, N. and Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10). doi:https://doi.org/10.1016/j.jiph.2021.08.010.

18. Tunca, S., Sezen, B. and Balcioglu, Y.S. (2023). Content and Sentiment Analysis of The New York Times Coronavirus (2019-nCOV) Articles with Natural Language Processing (NLP) and Leximancer. *Electronics*, [online] 12(9), p.1964. doi:https://doi.org/10.3390/electronics12091964.

19. Özsezer, G. and Mermer, G. (2022). Discussions About COVID-19 Vaccination on Twitter in Turkey: Sentiment Analysis. *Disaster Medicine and Public Health Preparedness*, pp.1–25. doi:https://doi.org/10.1017/dmp.2022.229.

20. Yan, C., Law, M., Nguyen, S., Cheung, J. and Kong, J. (2021). Comparing Public Sentiment Toward COVID-19 Vaccines Across Canadian Cities: Analysis of Comments on Reddit. *Journal of Medical Internet Research*, 23(9), p.e32685. doi:https://doi.org/10.2196/32685.

21. Verma, R., Chhabra, A. and Gupta, A. (2022). A statistical analysis of tweets on covid-19 vaccine hesitancy utilizing opinion mining: an Indian perspective. *Social Network Analysis and Mining*, 13(1). doi:https://doi.org/10.1007/s13278-022-01015-2.

22. coronavirus-yellowcard.mhra.gov.uk. (2022). *Homepage* | *Coronavirus (COVID-19)*. [online] Available at: https://coronavirus-yellowcard.mhra.gov.uk/.

Links for Python-related coding references:

1. Topic Modeling with Scikit Learn. Latent Dirichlet Allocation (LDA) is a… | by Aneesha Bakharia | ML Review

2. GitHub - LoLei/redditcleaner: Cleans Reddit Text Data :broom:

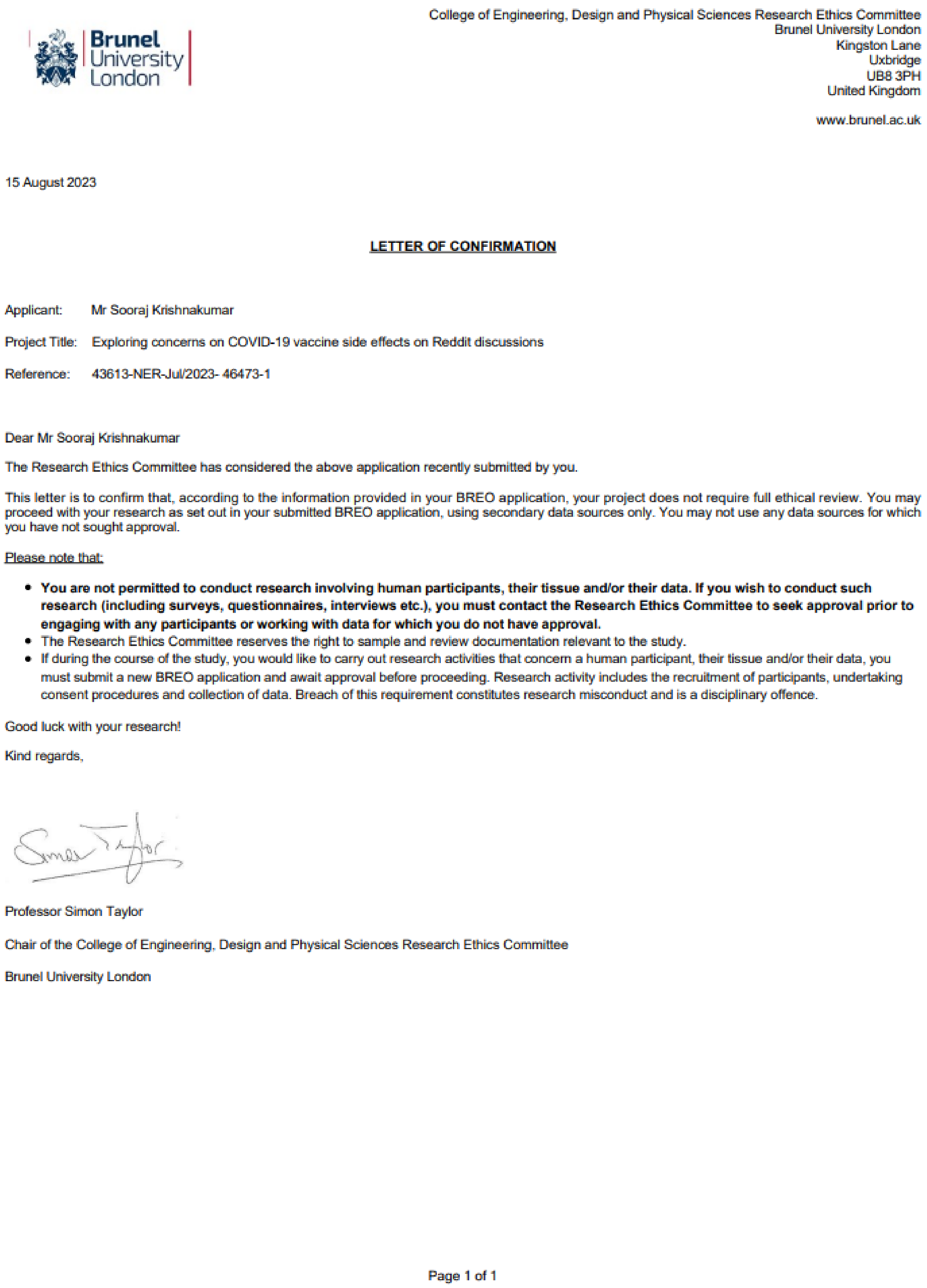
3. Evaluation Metrics for Classification Problems with Implementation in Python | by Venu Gopal Kadamba | Analytics Vidhya | Medium

4. NLTK vs SpaCy: Which Library Is Best for Named Entity Recognition? | by Dr. Soumen Atta, Ph.D. | Dev Genius

5. How do I remove words from my wordcloud? (Python 3) - Stack Overflow

6. Text Classification with Python and Scikit-Learn (stackabuse.com)

APPENDIX A : ETHICAL APPROVAL



Brunel University London

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

15 August 2023

LETTER OF CONFIRMATION

Applicant: Mr Sooraj Krishnakumar

Project Title: Exploring concerns on COVID-19 vaccine side effects on Reddit discussions

Reference: 43613-NER-Jul/2023- 46473-1

Dear Mr Sooraj Krishnakumar

The Research Ethics Committee has considered the above application recently submitted by you.

This letter is to confirm that, according to the information provided in your BREO application, your project does not require full ethical review. You may proceed with your research as set out in your submitted BREO application, using secondary data sources only. You may not use any data sources for which you have not sought approval.

Please note that:

- **You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research (including surveys, questionnaires, interviews etc.), you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.**
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must submit a new BREO application and await approval before proceeding. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,

Simon Taylor

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

Page 1 of 1

APPENDIX B: FILES DESCRIPTION

I have submitted 4 files out of which 1 is the Ethical Approval letter. Rest of the 3 are a CSV file and 2 Python code files (.ipynb).

1. Reddit_post_retrieval.ipynb => This file, as the name suggests isused to extract posts from the Reddit platform from different subreddits.

2. Corona1.csv => The CSV file obtained from the post retrieval code.

3. Reddit_Data_Analysis.ipynb => This file is the main code file where the Corona1.csv is used for data analysis purpose.

4. Letter.pdf => Ethical Approval Letter