

Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Information Technologies



Bachelor Thesis

**Analysis of Twitter content related to Covid-19
vaccination using NLP**

Pamela Vogli

© 2023 CZU Prague

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Pamela Vogli

Informatics

Thesis title

Analysis of Twitter Content Related to Covid-19 Vaccination Using NLP

Objectives of thesis

During the pandemic, all social media channels became a crucial source of information about COVID-19 for the public. The main objective of the thesis is to analyze the relationship between the Covid-19 vaccination rates and Twitter content.

Partial objectives:

- To identify relevant hashtags, download the Twitter dataset and conduct sentiment analysis.
- To build a model and run statistical analysis.
- To evaluate the results and interpret findings.

Methodology

The methodology of this thesis is based on reviewing and analyzing tweets about vaccines to determine and evaluate positive or negative sentiment within the text based on natural language processing (NLP). The dataset is retrieved from Twitter using a Python package called Tweepy. Following extraction, a statistical analysis is performed. Based on the findings, conclusions are conducted.

The proposed extent of the thesis

50 pages

Keywords

Vaccine, Covid-19, social media, hashtags, tweets, content analysis

Recommended information sources

BARUA, Ranjit; SUDIPTO, Datta; NIBEDITA, Bardhan. Impact of Social Media on Pandemic COVID-19 Outbreak. Handbook of Research on Representing Health and Medicine in Modern Media. IGI Global, 2021.

DUA, Sejal, Sentiment Analysis of COVID-19 Vaccine Tweets. Towards Data Science, 2021.

RESHI, Aijaz Ahmad, et al. COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset. MDPI, 2022.

SHAMRAT, F. M. J. M., et al. Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indones. J. Electr. Eng. Comput. Sci 23.1, 2021.

YOUSEFINAGHANI, Samira, et al. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. International Journal of Infectious Diseases, 2021.

Expected date of thesis defence

2022/23 SS – FEM

The Bachelor Thesis Supervisor

Ing. Miloš Ulman, Ph.D.

Supervising department

Department of Information Technologies

Electronic approval: 14. 7. 2022

doc. Ing. Jiří Vaněk, Ph.D.

Head of department

Electronic approval: 27. 10. 2022

doc. Ing. Tomáš Šubrt, Ph.D.

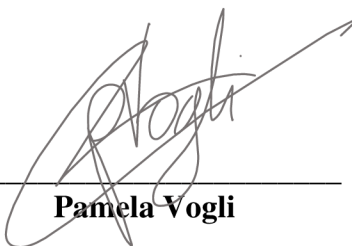
Dean

Prague on 30. 12. 2022

Declaration

I declare that I have worked on my bachelor thesis titled "Analysis of Twitter Content Related to Covid-19 Vaccination Using NLP" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on March 15, 2023



Pamela Vogli

Acknowledgment

There are many people I want to acknowledge who got me to this point in my life. But the most significant one is my sister, Xhesilda. You were my first critique partner. You gave me some of the harshest and best feedback I've ever received. Your ideas and thoughts are so prevalent in everything I do, I can hardly distinguish them from my own anymore. Most important of all, thank you for supporting me all these years.

Analysis of Twitter content related to Covid-19 vaccination using NLP

Abstract

The world was introduced to the COVID-19 virus in December 2019, which caused widespread concern and devastation. The development of a vaccine was seen as the most promising solution to the problem, leading to extensive research by organizations worldwide. However, the discovery of a significant number of people who were hesitant to take the vaccine prompted further investigation. This study aims to examine these hesitancy levels by analyzing tweets on Twitter. A sample of a substantial number of tweets was collected between January 2020 and January 2021. The tweets were subjected to natural language processing (NLP) based on sentiment and statistical analysis. The results provide valuable insights into the attitudes and opinions of people regarding COVID-19 vaccination. The study also highlights the importance of understanding the public's perception of vaccines and the need for effective communication strategies to address vaccine hesitancy.

In conclusion, the analysis of Twitter content related to COVID-19 vaccination provides a unique perspective on the public's perception of vaccines and contributes to the ongoing discussion about vaccine hesitancy.

Keywords: Vaccine, Covid-19, social media, Twitter, hashtags, tweets, content analysis, NLP

Table of content

1	Introduction	11
2	Objectives and Methodology	12
2.1	Objectives	12
2.2	Methodology	12
3	Literature Review	13
3.1	Pandemic	13
3.1.1	Pandemics throughout history	13
3.1.2	Coronavirus disease	15
3.2	Social media	15
3.2.1	Role of social media in COVID-19	17
3.2.2	Vaccine content on social media	19
3.3	Natural language processing	21
3.3.1	How does NLP work?	22
3.3.1.1	Natural language processing techniques	22
3.3.1.2	Levels of natural language processing	24
3.3.2	Common examples of NLP	26
3.3.3	Advantages and challenges	27
3.4	Related studies	29
3.5	Summary of main findings	31
3.6	Research question	32
4	Practical Part	33
4.1	Data preparation	33
4.1.1	Data retrieval and methodology	33
4.1.2	Cleaning data	35
4.2	Data evaluation	39
4.2.1	Approach 1: Sentiment Analysis	40
4.2.2	Approach 2: Statistical Analysis	44
5	Results and Discussion	51
5.1	Recommendation for future researchers	58
6	Conclusion	60
7	References	62
8	List of pictures, tables, graphs, and abbreviations	67
8.1	List of pictures	67
8.2	List of tables	67

8.3	List of graphs.....	67
8.4	List of abbreviations.....	67
	Appendix.....	69

1 Introduction

The World Health Organization has identified vaccine hesitancy as a significant threat to world health as a result of the recurrence of diseases that can be prevented by vaccination. There are a variety of information sources that may contribute to vaccine hesitancy, including the Internet and social media. As access to technology has increased, social media has become widespread across the globe. Through social media, anyone may quickly create and share content around the world without restriction. Users can also choose their own content streams. Therefore, anti-vaccination messaging on these platforms have raised a fair public health concern. Future vaccines, such as those designed to prevent COVID-19, could be compromised by this.

The third chapter discusses the current position of social media platforms in propagating vaccine hesitancy and introduces the method used to draw conclusions in the following chapters.

The *Practical Part* chapter provides an overview of the study and its objectives, with a focus on the sentiment analysis of Twitter data related to Covid-19 vaccination. It is divided into three sections. The first section, Related Works, provides a comprehensive discussion of the relevant studies in the field of sentiment analysis. The section on Data Preparation delves into the details of data collection and explains the methodology used to analyze the data. The final section of this chapter, Data Evaluation, presents two different approaches used in the analysis of the data and explains how these approaches were implemented to provide a comprehensive understanding of the tweets related to Covid-19 vaccination.

The fifth chapter of this thesis presents the results of the sentiment analysis of Twitter data related to Covid-19 vaccination. The limitations of the study are also discussed, and recommendations are provided for future researchers to further improve the analysis and increase the accuracy of the results.

The final chapter of the thesis provides a conclusion of the study and highlights the main findings. This chapter summarizes the results of the sentiment analysis of Twitter data related to Covid-19 vaccination and presents the implications of these findings for both the research community and the wider public.

2 Objectives and Methodology

2.1 Objectives

During the pandemic, all social media channels became a crucial source of information about COVID-19 for the public. The main objective of the thesis is to analyze the relationship between Covid-19 vaccination rates and Twitter content.

Partial objectives

- To identify relevant hashtags, download the Twitter dataset and conduct sentiment analysis.
- To build a model and run statistical analysis.
- To evaluate the results and interpret findings.

2.2 Methodology

Methodology of this thesis is based on reviewing and analyzing tweets about vaccines to determine and evaluate positive or negative sentiment within the text based on natural language processing (NLP). The dataset is retrieved from Twitter using a Python package called Tweepy. Following extraction, a statistical analysis is performed. Based on the findings, conclusions are drawn.

3 Literature Review

Social media is a great source to communicate information to the general public, especially during a pandemic. There is a certain correlation between emerging infectious diseases, such as COVID-19 and an increased usage of social media platforms by the public. As the information on social media is created and shared by users, this information is prone to be inaccurate or subjective, and often includes conspiracy theories and misinformation. Therefore, social media plays a crucial role in how individuals perceive their risk of getting exposed to a disease, the decisions that follows, and their risk-taking behaviors. This chapter will overview and analyze this relationship between social media and COVID-19 and identify the method used to draw conclusion on chapter four.

3.1 Pandemic

Throughout history, frequent outbreaks of infectious diseases have had a significant and long-lasting impact on society. Human civilization has been profoundly shaped by those events at every level, with their effects often lasting hundreds of years. There is no doubt that pandemic outbreaks decimated society, affected war outcomes, and led to the extinction of entire populations, but paradoxically their impact also helped foster innovations in many fields.

The scientific community was compelled to evolve the concepts of epidemiology, prevention, immunization, and antibiotic treatments as a result of epidemic outbreaks, which have helped define some of the fundamental foundations of modern medicine.

This chapter provides an overview of some of the most major epidemics that have ever occurred in human history. By identifying the extent of the effects of a pandemic, we can better understand the social, medical, and psychological impacts these events had on our civilization.

3.1.1 Pandemics throughout history

In an age of civilization, when humans-built cities, and trade routes, and fought with one another, pandemics became more likely. Table 1 shows how pandemics changed history through their devastating effects on human populations. The death tolls of each pandemic are also visualized as a percentage of the total global population at that time (LePan and Schell 2022).

NAME	TIME PERIOD	TYPE / PRE-HUMAN HOST	DEATH TOLL	GLOBAL POPULATION LOST
Antonine Plague	165-180	Believed to be either smallpox or measles	5M	2.6%
Japanese smallpox epidemic	735-737	Variola major virus	1M	0.5%
Plague of Justinian	541-542	Yersinia pestis bacteria / Rats, fleas	30-50M	19.1%
Black Death	1347-1351	Yersinia pestis bacteria / Rats, fleas	200M	51%
New World Smallpox Outbreak	1520 – onwards	Variola major virus	56M	12.1%
Great Plague of London	1665	Yersinia pestis bacteria / Rats, fleas	100,000	0.02%
Italian plague	1629-1631	Yersinia pestis bacteria / Rats, fleas	1M	0.2%
Cholera Pandemics 1-6	1817-1923	V. cholerae bacteria	1M+	0.1%
Third Plague	1885	Yersinia pestis bacteria / Rats, fleas	12M (China and India)	1%
Yellow Fever	Late 1800s	Virus / Mosquitoes	100,000-150,000 (U.S.)	0.015%
Russian Flu	1889-1890	Believed to be H2N2 (avian origin)	1M	0.08%
Spanish Flu	1918-1919	H1N1 virus / Pigs	40-50M	2.5%
Asian Flu	1957-1958	H2N2 virus	1.1M	0.03%
Hong Kong Flu	1968-1970	H3N2 virus	1M	0.02%
HIV/AIDS	1981-Present	Virus / Chimpanzees	25-35M	0.7%
Swine Flu	2009-2010	H1N1 virus / Pigs	200,000	0.002%
SARS	2002-2003	Coronavirus / Bats, Civets	770	>1.22E-05 %
Ebola	2014-2016	Ebolavirus / Wild animals	11,000	>1.5E-04 %
MERS	2015-Present	Coronavirus / Bats, camels	850	>1.16E-05 %
COVID-19	2019-Present	Coronavirus – Unknown (possibly pangolins)	6.3M (Johns Hopkins University estimate as of July 8, 2022)	0.08%

Table 1 - Pandemics timeline (LePan and Schell 2022)

Five of the world's worst pandemics according to their death tolls are: Black death (200M), Smallpox (56M), Spanish Flu (40-50M), Plague of Justinian (30-50M), and last but not least HIV/AIDS (25-35M). In contrast to prior pandemics, Covid-19 would be ranked third from the bottom. One of the most surprising events so far is how little nations did to properly prepare for it, despite months of seeing it unfold in China.

3.1.2 Coronavirus disease

Coronaviruses have been known for more than 50 years. The term "corona" has a variety of connotations, but virologists decided the name coronaviruses by drawing a comparison between the solar corona and the characteristic projections on the outside of the virus. In the majority of cases, this virus causes severe acute respiratory syndrome, which is fatal. Coronaviruses are single-stranded RNA viruses with a diameter of roughly 120 nanometers. Mutation is a frequent occurrence in them, making them a highly diverse group of organisms (Aronson 2020).

Before the outbreaks of SARS-COV-1 in 2003, MERS in 2012, and SARS-COV-2 in 2019 that resulted to significant human infections, the first coronaviruses discovered to infect people were known as 229E and OC43. However, they only caused extremely minor infections.

The COVID-19 virus is believed to have started in bats and subsequently transmitted to people through the meat sold in China's meat markets, which were contaminated with the feces of wild animals (Anderson, et al. 2004).

3.2 Social media

A social media platform is an internet-based application that allows users to collaborate, share, create, and interact with each other through a variety of content types available. The overall response is usually greater, during major events, whether they're sports, diseases, or natural disasters. A solid illustration of how social media platform searches have been shown to be correlated with diseases is how information searches on these platforms peaked in China 10–14 days before the outbreak of COVID-19 cases (Abd-Alrazaq, et al. 2020) (Cuilian , et al. 2020).

Social media platforms have also positively impacted the public by allowing them to stay in touch with friends and family to minimize loneliness and boredom, which have been linked to

anxiety and long-term distress, making them a crucial guideline for home isolation to help ease the psychological impact (Brooks, et al. 2020).

One of the most important qualities of social media platforms in this pandemic has been the rapid transmission of protocols at every levels. Now, it's a common practice to share treatment protocols, safety equipment, and even suggestions for fair resource distribution in medical field (Emanuel, et al. 2020). This makes it possible for centers with limited resources to quickly design standards that can be implemented or modified to fit their unique circumstances or resources. Twenty years ago, before the majority of social media channels existed, this would have been unimaginable.

The following is a summary of the main benefits and drawbacks of using social media during the pandemic.

Advantages of social media use

In the era of COVID-19, social media have been extremely useful in the rapid dissemination of educational content, for example, creating an infographic which describes how to manage patients with confirmed or suspected COVID-19. After being shared on Twitter and WeChat, it was translated into more than ten languages within a few days. Furthermore, its distribution provided a unique opportunity to adapt the infographic to the specificities of each healthcare setting (Nickson, et al. 2020).

It was beneficial to disseminate information regarding preventive measures more quickly. According to a recent study by Basch, et al. 2020, the top 100 videos on YouTube with the keyword "coronavirus" altogether had more than 165 million views, and 85% of them being news channels; It was found that less than one-third of the videos included the suggested preventative actions, and only half highlighted the most common symptoms, while approximately 90% of them made comments on fatalities, anxiety, and the quarantine situation. This study reminds everyone to think critically about the chances missed to share reliable information on COVID-19's typical symptoms and transmission prevention on websites like YouTube, which are widely used as sources of information (Basch, et al. 2020).

Studies on publications have revealed a rise in downloads, searches, and cites of these articles: (Teoh, et al. 2020, Allen, et al. 2013) when they are shared on social media platforms such as Facebook, Twitter, etc. In addition to significantly reducing editing times, which have gone from months of processing to days or weeks from its receipt, these qualities have surely enabled for the quick transmission of knowledge globally during the COVID-19 epidemic.

The ability to organize collaborative scientific studies, questionnaires, and multi-center research has been another benefit of social media platforms during the COVID-19 pandemic. Lastly, social media platforms have the advantage of promoting ongoing medical education using live or recorded conferences via platforms like Zoom, YouTube, or Skype (Nickson, et al. 2020).

Disadvantages of social media use

One of the drawbacks is the possibility for the information shared to be outdated, invalid, information that hasn't been critical analysis nor it doesn't apply to our surroundings or is simply fake. The "bubble filters," a term introduced by Eli Pariser in 2011, is another serious obstacle of social media use. The study describes a "personalized ecosystem" aimed towards the user, in which algorithms use data gathered from the same user to predict their preferences and produce results that are thought to be similar to the interests of that user. (Pariser 2011). These bubbles create a cycle of repetitive information that prevents the user from ever seeing alternative sources from which they can compare information. This theory applies to any sickness or condition that is looked up online or even on social media sites such as Twitter and Facebook (Holone 2016).

The potential for spreading false, alarming, and exaggerated information that can provoke fear, anxiety, and depression in individuals with or without preexisting mental issues is perhaps the worst aspect of social media. In a study by Wang et al. 2020 in China, who conducted an online survey with 1,210 respondents, they found that 53.8% admitted that the epidemic had a moderate or severe psychological impact (Cuiyan Wang 2020); moreover in an effort to assess the level of stress and anxiety in the population and establish appropriate measures to prevent sequels, a research group developed a scale titled "Fear of COVID-19 scale". It was found out that post-traumatic stress disorder (PTSD) is one of the most common psychiatric sequelae that resulted from the SARS outbreak in Asia in 2003, followed by depression (Ahorsu, et al. 2022, Mak, et al. 2009). In some parts of the world, such as India, Britain, Germany, and Italy, more severe diseases and events are already reported, including suicides (Jain and Thakur 2020).

3.2.1 Role of social media in COVID-19

Social isolation measures resulted in an increase in the usage of social media platforms for information sharing (Abd-Alrazaq, et al. 2020). Based on 2021 data, social media usage climbed by 11% over the previous year, reaching 4.2 billion users (54%) (Kemp 2021). In

circumstances like emergencies and natural catastrophes, where quick information flow is essential, it has been shown that the use of social media platforms increases rapidly. Given how simple it is to access the latest information on social media, using it in times of health crisis is seen as a good opportunity (Denecke, et al. 2013, Guess, et al. 2020, Barua, Datta and Bardhan 2021).

For instance, there was an increase in social media use and compliance with safety precautions including handwashing, face mask use, and social distancing during the SARS, Ebola, and H1N1 outbreaks (M. Fast, Markuzon and Louis 2019, Collinson and M Heffernan 2014, Sun, Zhu and Cui 2007). Similar to this, social media helped to spread quickly, the COVID-19 protective equipment, social isolation, and even medical treatment protocols. Information on COVID-19 was posted on the websites of international organizations like the WHO (World Health Organization) and the Center for Disease Control, and it was made available for use even in the most distant regions of the world. Contrary, social media has also promoted the unmonitored spread of inaccurate information about the pandemic, as well as unsupported claims regarding the virus's origin, diagnosis, prevention, and treatment, as well as conspiracy theories and scientifically unsupported cures (Bridgman, A, et al. 2020, Jargin 2010, Tasnim, Hossain and Mazumder 2020, Brennen, JS, et al. 2020).

The term "infodemic," which is defined by WHO as "too much information including false or misleading information in digital and physical environments during a disease outbreak," came to the forefront with COVID-19 (World Health Organization n.d.).

Conspiracy theories and disinformation about vaccines have been spread on social media as a result of treatment choices and the launch of vaccines. (Salathé, et al. 2012) The amount of specialist commentary on the anti-vaccine movement, conspiracies, and erroneous assumptions regarding the pace of vaccine development, long-term side effects, and challenges with the COVID-19 vaccine increased across the media.

Social media and the internet are incredibly successful in spreading knowledge about the security and effectiveness of vaccines. However, once people have been exposed to inaccurate information, it can be challenging to reverse their negative attitudes against vaccinations (Kata 2010, Del Vicario, et al. 2016, Eysenbach and Chew 2010). It is quite likely that parents who use the internet to research vaccines, hold the opinion that healthy children should not get vaccinated and that the vaccine does more harm than good (Jones, et al. 2012).

In order to understand and control public perceptions of risks to society during the COVID-19 pandemic, it is extremely valuable to monitor social media debates (J. Barnett, Alperstein and Pascual-Ferrá 2020). Social media exposes people to a constantly increasing amount of information because it allows users to write and share information. But it also contributes to the spread of misinformation by uniting those who share the same beliefs and giving them a sense of authority (Guess, et al. 2020, Smith, Himmelboim and McCreery 2013, Karafillakis, et al. 2021).

SNA (Social network analysis) is an analytical technique used to analyze the structure of social media and the interactions among participants (O'Donohoe 2015). SNA has been implemented regularly by researchers for a variety of studies using multiple data sources.

One of the most powerful and influential social networks is Twitter. The fact that all content is public, makes it possible for researchers to gather data without obtaining permission from users. Twitter is a microblogging site where users may post and engage with each other about various topics. They interact with each other by using hashtags, retweets, and replies, as well as following different accounts. Twitter is a great sharing information tool.

Finding influencers is one of SNA's goals. On Twitter, influence is defined as "the potential of a user's activity to inspire a subsequent action by another user." To measure influence there are some indicators of influence such as comments, retweets, and tags (Leavitt, et al. 2009).

The number of studies that links social media and vaccines has increased globally. The next chapter will discuss some of these studies and will determine how the community interacted with vaccinations and gain a better understanding of vaccine-related tweets.

3.2.2 Vaccine content on social media

Several studies have examined how vaccine content is portrayed on social media platforms and more broadly on the Internet (Signorelli, et al. 2015, Arif, et al. 2018, Chapman, Davies and Leask 2002, Gunaratne, Haghbayan and Coomesb 2019, Basch and MacLean 2019, Moreno, et al. 2019).

In 2017, Basch et al. reviewed 87 YouTube videos that had the terms "vaccine safety" and "vaccines and children." Among these, 65% stated objection to vaccinations, yet only 5.6% were created by government experts, and 36.8% lacked any supporting scientific data. Additionally, 27.5% of the top YouTube videos found by searching "COVID-19" and

"coronavirus" had inaccurate material and had already received more than 60 million views (Basch, et al. 2017, Chan, et al. 2020).

Studies show that anti-vaccine content increases user interaction. For instance, Blankenship et al. 2018 found that anti-vaccine tweets were more likely to be retweeted than neutral tweets when they analyzed tweets including the "vaccine" hashtag (#vaccine) between 2010 to 2016 (Blankenship, et al. 2018). Similar findings were made by Basch et al. when they examined 150 Instagram posts with the hashtag #HPV and discovered that anti-vaccine posts had a significantly greater number of likes (Basch, et al. 2017).

Social media discussions about vaccination have changed over time, with trends frequently reflecting actual events. Gunaratne et al. 2019 showed that the measles outbreak in 2014–2015, the publishing of the anti-vaccine book *Vaccine Whistleblower*, and the film *Vaxxed*, coincided with a significant increase in an anti-vaccine speech on Twitter in 2015. This study also showed that pro- and anti-vaccine information may naturally distinct into separate groups, presumably as a result of self-selection on social media that combines communities with similar viewpoints. While pro-vaccine content on Twitter mainly revolved around the hashtag #vaccineswork, anti-vaccine content mainly developed a community focused around #cdcwhistleblower and #vaxxed hashtags (Gunaratne, Haghbayan and Coomesb 2019, Ortiz-Ospina 2019). Content appears to be shared more frequently between users who have similar views on vaccination than it does between those who maintain opposing views, indicating that while the design of such platforms may give the appearance of debate, in reality, it mostly serves to support long-held beliefs rather than considering new ones. Such ideological segregation may prevent public health to promote on social media vaccination (Yuan, Schuchard and Crooks 2019).

Since COVID-19 spread rapidly and resulted in a global pandemic, social media discourse has centered around COVID-19. Twitter reports a COVID-19-related tweet every 45 milliseconds, making #coronavirus the second most common hashtag for 2020 (Lambe and Josephson 2020, Cinelli, et al. 2020). On social media sites, inaccurate and misleading information regarding COVID-19, possibly harmful treatments, and vaccination are unfortunately still spreading (Kouzy, et al. 2020, Brennen, JS, et al. 2020). In the early stages of COVID-19, Cinelli et al. investigated this "infodemic," by applying epidemic modelling to the transmission of information on multiple social media platforms to determine the "transmissibility" of posts on

every platform. Regardless of the platform, there weren't any noticeable changes between spreading patterns of trustworthy information and doubtful ones (Cinelli, et al. 2020).

Furthermore, there is a fair concern that discussions on vaccines may not only be based on real human accounts. In their study, Broniatowski et al. 2018 identified two categories of artificial users: "bots" and "trolls". Bots accounts generate content automatically and trolls falsely claim to be someone else. There was a tendency for content-polluting bots to increase anti-vaccination content, whereas troll accounts were more likely to increase both, pro- and anti-vaccine content, resulting in a false sense of equilibrium regarding vaccination safety and benefits (Broniatowski, et al. 2018).

Ferrara examined the content and spreading patterns of more than 62 million tweets using a mix of machine learning and manual verification. He found that, compared to non-bot accounts, automated bot accounts posted significantly more tweets about COVID-19. Also, the content analysis of these tweets revealed that, in addition to the COVID-19 content that was posted, these tweets also highly promoted political conspiracies and divisive hashtags (Ferrara 2020). Such conspiracies can have harmful consequences, such as the spread of rumors associating 5G cellular networks with the spread of COVID-19, which led to cell tower vandalism (Wasim, et al. 2020).

Overall, these studies indicate that while the proportion of pro- and anti-vaccine content depends on the platform, the amount of user interaction generated by anti-vaccine content is frequently higher than that of pro-vaccine content and that this user engagement may be influenced by the occurrence of relevant current events and the activity of bots and trolls accounts.

3.3 Natural language processing

Definition of Natural language processing is not an easy task because it is a highly active area. Although there isn't currently a universally agreed-upon definition, there are some aspects that are certain to be included in any knowledgeable definition.

Natural language processing (NLP) is an artificial intelligence subfield that supports computers in understanding, interpreting, and manipulating human language. The goal of NLP is to fill the gap between human communication and computer understanding through the use of a variety of

disciplines, including computer science and computational linguistics (Reshamwala, Mishra and Pawar 2013, Kumar 2011, Shamrat, et al. 2021). Illustration of a natural language interface:

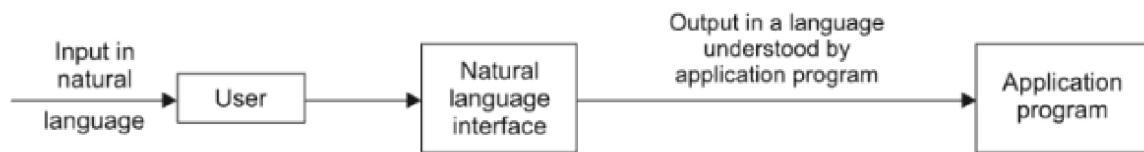


Figure 1 - Block diagram of an NLP system (Kumar 2011)

3.3.1 How does NLP work?

Machines must be able to recognize, the rules of a natural language, somewhat. Therefore, NLP deals with different aspects of language: phonology, morphology, syntax, semantics, and pragmatics. And its worst enemy is called ambiguity. Each one of these aspects adds another layer of contextual understanding of words. But in most cases, the language we are aiming to process must be first transformed into a structure that the computer is able to read. This is no different than making a child learn how to read for the very first time. In order to clean up a dataset and make it easier to interpret, different techniques are used to achieve the purpose of NLP (Ribeiro 2021, Couto 2015).

3.3.1.1 Natural language processing techniques

As stated above, NLP techniques involve various linguistic levels to understand and use language. The linguistic levels of NLP will get discussed in the next subchapter. Using Natural Language Processing techniques, computers can now recognize and interpret the meaning behind words, phrases, and paragraphs accurately (Krishna, Akila and Rajesh 2019).

Let's take a closer look at some of the techniques used in NLP in practice:

- *Tokenization*

Tokenization, a crucial step in natural language processing, is the process of dividing a string of words into tokens, or units of meaning. This will make it easier for the program to comprehend both the single words and how they fit into the overall text. This is crucial for bigger amounts of text since it enables the machine to count the frequency of specific words and where they were used. (e.g., New York) (Horan 2020).

- *Stop words*

In NLP text processing, eliminating stop words is a crucial step. It entails removing high-frequency words like which, to, at, for, and is, which have little to no semantic meaning for a sentence. Additionally, this method is used to modify lists of stop words to contain words that must be ignored during NLP text processing (Hvitfeldt and Sigle 2021).

- *Lemmatization and Stemming*

Inflected forms of words (words in their different grammatical forms) are part of everyday language. Lemmatization and stemming are two techniques that NLP use to return these words to their root form and make them simpler for computers to comprehend. A lemma is the term used to describe the word in its basic form. The words "is," "are," "am," "were," and "been," for instance, are grouped together under the lemma "be."

However, the root form of a word is called a stem when we refer to stemming. Stemming "trims" words, therefore, word stems might not always be semantically appropriate. For example, the root form "consult" would result from stemming the words such as "consultant," "consulting," and "consultants."

While lemmatization uses dictionaries and selects the right lemma depending on context, stemming uses isolated words without taking the context into account. For example, in the sentence: "This is better". Lemmatizers alter the word "better" into the word "good," whereas stemming maintains the same word and leaves it unchanged. Stemmers are quicker and simpler to build than lemmatizers, but they can produce findings that are less precise, therefore lemmatizers are recommended for more precise linguistic rules (Hvitfeldt and Sigle 2021).

- *Part-of-speech Tagging*

Part-of-speech tagging (abbreviated as PoS tagging) entails giving each token within a text a part of speech category. Verb, adjective, noun, pronoun, conjunction, preposition, and intersection are a few examples of popular PoS tags. PoS tagging is useful for understanding phrase meaning by determining links between words (Chiche and Yitagesu 2022).

- *Named Entity Recognition*

One of the most common tasks in the semantic analysis is named entity recognition, which entails identifying entities within a text. Entities include things like names (E.g., Audrey Hepburn), locations (E.g., Machu Picchu), companies (E.g., Google), email addresses, and more. NER has a two-step process: It detects a named entity and then it categorizes it. This technique can be used in human resources to categorize CVs or to improve response time for customer support, by filtering complaints using priority keywords (Gandhi 2022).

3.3.1.2 Levels of natural language processing

The "levels of language" approach is the clearest way to explain what actually is happening in a Natural Language Processing system. As one of the necessities for any NLP task, natural language understanding is also one of the most challenging one, since it's difficult for a computer to process thoughts, have world knowledge and represent linguistic input at the same time.

In most NLP systems, the first step is determining morphological structure before moving on to more advanced analysis such as grammar, word order, and meaning (Kumar 2011).

The process of understanding and extracting meaning from a text or spoken word is divided into seven interdependent levels. The seven-level approach follows as below (Kumar 2011, Khurana, et al. 2022, Liddy 2001):

- **Phonology:** is the interpretation of speech sounds within and across words. As part of an NLP system that accepts spoken input, the sound waves are analyzed and encapsulated into digital signals that can be interpreted based on rules or by comparing them with language models.
- **Morphology:** conceptual analysis of words, focusing on roots, prefixes, and suffixes. Example: The word 'precancellation' can be morphologically scrutinized into three separate morphemes: pre (the prefix) – cancel (the root) -tion (the suffix). Each morpheme has the same meaning across words, so humans can decode an unknown

word by breaking it down into its constituent morphemes. In the same way, an NLP system can interpret each morpheme to gain its meaning.

- **Lexical:** this stage of natural language analysis involves evaluating the lexical accuracy of words. Lexicon stands for dictionary. It is a group of all grammatically correct words, along with their definitions. In NLP, the first stage of input text processing involves scanning each word in the phrase and searching for all the significant linguistic data about that word. The rules and information required for performing the first stage analysis are provided by the lexicon.
- **Syntactic:** this level focuses on dissecting the words of a sentence to determine the phrase's grammatical structure. Both a grammar and a parser are needed for this. The result of this level is a projection of the sentence that shows the relationships between the words' structural dependencies. There are numerous grammars that can be used, each of which will influence the parser that is used. Most languages utilize syntax to express meaning because order and dependency are important components of meaning. It may seem that the two sentences 'The dog chased the cat.' and 'The cat chased the dog.' differ only in syntax, but their meanings are quite distinct.
- **Semantic:** determining a sentence's potential interpretations, including word meanings based on context. For example, amongst other meanings, 'match' can mean either a game of sport, or a small wooden stick used for lighting fires (Twinkl n.d.).
- **Discourse:** Interpreting larger-scale texts to understand their structure and meaning. It analyzes logical structure by connecting different words and sentences that assure coherence. It focuses on the components of the text that bring out the meaning through analyzing the correlations between sentences and uncovering linguistic structures from texts of various levels.
- **Pragmatic:** an understanding of how language is used in a purposeful manner in situations, particularly when world knowledge is required. It deals with what the listener and the speaker assume from each other. As a matter of fact, it analyzes non-

directly spoken sentences. A meaningful interpretation of the text is derived by analyzing the context.

3.3.2 Common examples of NLP

NLP has a wide variety of everyday uses that are both common and useful. Here are a few additional examples apart from virtual assistants like Alexa or Siri:

- **Email filters**

One of the most popular and fundamental applications of NLP is in email filters. Although they weren't totally accurate when they were initially launched, emails now rarely go up in the wrong inbox because of decades of machine learning training on millions of data samples. For example, Gmail's email classification is one of the more common and recent NLP applications. Based on their content, the system determines whether emails fall into the primary, social, or promotional category (Tableau Software, LLC n.d., MonkeyLearn Inc. n.d.)

- **Predictive text and autocorrect**

NLP is used every time a text message is sent through a smartphone. The texting app will often propose the right term after the user has only typed the first few letters of a word. And since NLP also picks up on user's messages, the more the user texts, the more accurate it gets. The more the user uses predictive text, the more it will adapt to user's unique linguistic preferences. (Wonderflow 2018, Tableau Software, LLC n.d.)

- **Machine translation**

One of the key areas of interest in the NLP research has always been the potential for translating speech and text into multiple languages. The use of NLP by online translators allows them to translate languages more accurately and provide grammatically correct results. One of the most recent developments in MT relates to software that can translate text and identify the language from inputted text. (MonkeyLearn Inc. n.d., Tableau Software, LLC n.d.)

- **Chatbots**

Although chatbots are not new, advances in NLP have made them more useful, to the point where some clients are no longer required to initiate contact with an actual representative.

Chatbots can assist customers in navigating articles, ordering goods, managing accounts, recommending services, providing discounts, and making bookings (MonkeyLearn Inc. n.d.).

- **Monitor brand sentiment on social media**

Sentiment analysis is a process that automatically categorizes opinions in a text as positive, negative, or neutral. It is frequently used to keep track of social media sentiment. Sentiment analysis can monitor and assess the tone of comments made about the overall brand, a specific product or feature, or even compare it to competition. Users can promptly respond to these negative comments by monitoring sentiment analysis (MonkeyLearn Inc. n.d.).

3.3.3 Advantages and challenges

The implementation of NLP certainly has its advantages and challenges, as summarized in Table 2.

Advantages

There are several advantages when using NLP, but here are a few of them (Wolff 2020, Kubicle 2022):

- *In-depth Data Analysis*
NLP helps machines to automatically read and analyze high volumes of unstructured text data, such as customer service complaints, social media comments, online reviews, news, and other types of data.
- *Streamlines Sophisticated Processes*
Machines can learn to sort and route information with the help of NLP tools, in a very precise and effective way, with little or no human interaction.
- *Enhances Customer Experience*
AI systems are capable of quickly extracting data, which helps in delivering responses to consumers on time. Chatbots that have been programmed to respond to common client inquiries, such as shipping charges, open hours, or other information about the service.
- *Better understand the market*
NLP algorithms may be customized to meet the demands of a business, including complicated industry-specific technology.

Challenges

The majority of the challenges come from data complexity. NLP is still a relatively new field of study, therefore there is a lot of room for engineers and companies to address the issues that arise with implementing NLP systems.

Let's take a look at some of those challenges below (Rosoka 2020, Kaushik 2021, Kubicle 2022).

- *Ambiguity*
Human language is ambiguous by nature, which is why NLP is so challenging. Even humans have difficulty analyzing and categorizing human language correctly. Sentences and phrases that might have two or more interpretations are referred to as ambiguous in NLP. However, one of the NLP techniques, part-of-speech (POS) tagging can make a small difference in solving this problem.
- *Training data*
In order to better understand a language, NLP focuses on studying it. Even the most advanced AI needs to spend a significant amount of time reading, listening to, and using the language in order to become fluent. This is true for everyone who wants to learn a new language well. An NLP system's capabilities are determined on the training data that receives. If the system is given skewed or inaccurate data, it will either learn the wrong things or learn inefficiently.
- *Development time*
To properly train an AI, millions of data points must be evaluated; processing all of that data on a slow PC may take a lifetime.
- *Phrases with Multiple Intentions*
The author or speaker may employ many styles to express the same topic, depending on their personality, intentions, and emotions. Some of them, like irony or sarcasm, can signify something completely different from what it means literally. Even though sentiment analysis has advanced significantly in recent years, accurately deciphering the pragmatics of the text is still a challenge.

Advantages	Challenges
In-depth Data Analysis	Ambiguity
Streamlines Sophisticated Processes	Training Data
Enhances Customer Experience	Development Time
Better understand the market	Phrases with Multiple Intentions

Table 2 - Advantages and Challenges of NLP

While Natural Language Processing has limits, it nonetheless provides enormous and far-reaching benefits to any organization. Many of these limitations will be overcome in the upcoming years as new methods and technologies emerge on a daily basis.

3.4 Related studies

Sentiment Analysis is an ongoing field of study in text mining. It involves the use of computerized methods to analyze the emotional tone and subjectivity of text. The goal is to understand people's opinions, feelings, and sentiments expressed in written content. In recent years, there has been a significant increase in research focused on sentiment analysis in medical blogs. The following sections will present a summary of previous work in this area (Monkey Learn n.d.).

The first study (Hussain, et al. 2021) utilized Naive Bayes, Decision Trees, and Random Forest algorithms to conduct sentiment analysis on social media data. The sentiments were classified as positive, negative, and neutral based on Twitter data. The results showed that Indonesian Twitter users tended to have a neutral stance. The accuracy of the three methods was compared using fast miner tools, with Naive Bayes having the highest accuracy of 86.43%. The accuracy of Decision Trees and Random Forest was 82.91%.

The second study (Prasad, Reddy and Puneeth 2020) examined the use of Naive Bayes Classifier and Decision Tree Algorithms for classification. It was suggested that these methods are more accurate and effective for predicting the type of brain tumor. The study aimed to show that the Decision Tree algorithm was more straightforward and simpler than the Naive Bayes algorithm. The two algorithms were used to identify the type of tumor and allowed for the examination of historical data from data sets to help neurologists make informed guesses. The

results showed that the Decision Tree algorithm was faster and more accurate than the Naive Bayes algorithm.

The third study compared Naive Bayes, Decision Trees, and Neural Networks in their ability to classify web pages. The findings of this study (Shi, et al. 2019) showed that the improved Naive Bayes classifier outperformed both the traditional Naive Bayes and Decision Trees algorithms.

In the fourth study, Decision Tree, K-Nearest Neighbor, and Nave Bayes were used to analyze sentiment from tweets. Text mining, preprocessing, categorization, and other techniques were used to develop sentiment analysis classification and analysis, with Rapid Miner also used to compare the analysis of attitudes using three different categories within the dataset. The findings of this study (Yousefinaghani, et al. 2021, Moreno-García, De la Prieta and Dang 2020) showed that the Nave Bayes algorithm had the highest accuracy of 77 percent.

(Yousefinaghani, et al. 2021) created a Twitter content-based analysis model to determine the popular thoughts and attitudes towards COVID-19 vaccinations. They analyzed retweets to uncover the prevailing opinions about vaccinations, including their growth over time, geographical location, main themes, hashtags, and account features. A supervised K-Nearest Neighbor (KNN) clustering algorithm was used to classify the data into positive, negative, and neutral categories.

(Khakharia, Shah and Gupta 2021) used Machine Learning to analyze public tweets related to COVID-19 vaccinations on Twitter to determine users' perspectives on the vaccine. The tweets were analyzed as positive or negative messages using three different machine learning algorithms, Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), and Logistic Regression (LR). The LR algorithm was found to be the most accurate with a 97.3 percent accuracy, followed by SVM with a 96.26 percent accuracy. On the other hand, MNB had the lowest accuracy of 88 percent.

Marcec and Likic (Marcec and Likic 2021) also conducted a study on vaccine drive analysis using Twitter-based sentiment analysis, suggesting that social media platforms like Twitter can be a useful tool to immediately observe people's feelings and beliefs on SARS-CoV-2 vaccinations.

In reference to (Khan, Alam and Dhruva 2021), they employed the use of VADER, a natural language processing tool known as Valence Aware Dictionary for sentiment Reasoning, to gauge individuals' emotions towards specific vaccines. They classified the gathered opinions into three categories - positive, negative, and neutral. The results showed that 33.96 percent of the participants had positive attitudes, 17.55 percent had negative attitudes, and 48.49 percent held neutral attitudes.

Meanwhile, (Nezhad and Deihimi 2022) conducted a study on the views of Iranians towards COVID-19 vaccinations by analyzing Twitter posts from Iran. They aimed to investigate the different attitudes of Iranians towards both domestically produced and imported COVID-19 vaccines. They used a CNN-LSTM architecture to determine the emotions expressed in the recovered tweets and to understand the general Iranian attitudes towards the COVID-19 vaccine.

In conclusion, this thesis was based on Twitter content and VADER was used for sentiment analysis because both have proven from above studies to be effective tools in understanding public perception and sentiment towards a particular topic. Twitter has become a prominent platform for discussion and the sharing of opinions, making it a valuable source for sentiment analysis. Additionally, the use of VADER as a sentiment analysis tool provided accurate and efficient results, demonstrating its usefulness in analyzing social media content.

3.5 Summary of main findings

This chapter is written with two objectives. The first objective gives insights of the persistent rise in vaccine-preventable diseases prompted the WHO to name vaccine hesitancy as a significant global health threat. Social media platforms and the Internet are among the sources of health information available to patients in the digital age. The popularity of social media platforms is leading to public health concerns regarding the potential impact of anti-vaccination content. Many studies indicate that while the proportion of pro- and anti-vaccine content depends on the platform, the amount of user interaction generated by anti-vaccine content is frequently higher than that of pro-vaccine content.

The second objective of this paper focuses on explaining theoretically how Natural Language Processing works as NLP is the methodology used in the practical part. To summarize shortly, the NLP field is part of artificial intelligence, where a variety of research backgrounds is used

to build models based on conventional machine learning algorithms, NLP techniques, and deep learning to extract information and understand documents. NLP is widely used in many fields due to the development of big data, deep learning, and cloud computing. NLP effectively improves efficiency and accuracy of work, which improves people's lives. NLP provides scholars with increased analytical accuracy for textual data as it is continuously improving and breaking through in algorithms and datasets. In the future, it is undeniable that NLP will have a positive impact on the management literature in the future, both theoretically and practically.

3.6 Research question

Social media's ascent has been accompanied by a sharp increase in the anti-vaccine content among platforms. Significant academic attention has been given to this correlation: Numerous studies on YouTube usage have been carried out (Basch, et al. 2017, Basch, et al. 2020, Chan, et al. 2020, Moreno, et al. 2019). These studies have consistently found that the anti-vaccine content is frequently higher than that of pro-vaccine content.

Despite this, little in-depth research has been done on Twitter. This paper sets out to address this research gap. Aiming to answer one research questions, a correlation between the Covid-19 vaccination and Twitter content. Based on the research question, this study reviewed and analyzed tweets about vaccines to determine and evaluate positive or negative sentiment within the text using natural language processing (NLP).

4 Practical Part

The practical part of this thesis aims to analyze Twitter content related to Covid-19 vaccination using NLP techniques. This chapter is divided into two main subchapters: Data preparation and Data evaluation. The first subchapter focuses on the data collection methodology, data cleaning, and data pre-processing, which are crucial steps for preparing the data for analysis. The second subchapter presents the data evaluation, which includes two approaches - Sentiment analysis and Statistical analysis - to provide a comprehensive understanding of the tweets related to Covid-19 vaccination.

4.1 Data preparation

Data preparation is a crucial step in the analysis process and involves transforming raw data into a format that can be effectively utilized for analysis. The process of data preparation involves cleaning, labelling, and transforming the data into a structured format. This chapter will provide an overview of the data preparation process that was carried out for the COVID-19 vaccination tweet analysis. The steps taken to collect, clean, and format the data will be discussed in detail, including the use of the Python library, Tweepy, for data extraction, and the use of natural language processing (NLP) techniques for analysis. Additionally, this chapter will also provide an overview of the sample size and the filters applied to ensure the quality of the data. The end result of this data preparation process will be a clean and structured dataset that will be used for further analysis in later sub-chapters.

4.1.1 Data retrieval and methodology

This sub-chapter outlines the process of extracting tweets related to COVID-19 vaccination from Twitter.

The extraction of tweets related to COVID-19 vaccination involved the utilization of the Tweepy Python library. Initially, a Twitter developer account was established, and API keys and tokens were obtained. These authentication credentials were then utilized to establish a connection to the Twitter API through the *tweepy.OAuthHandler* method (Dua 2021).

```
import tweepy  
# Enter your Twitter API keys and tokens
```

```

consumer_key = "your_consumer_key"
consumer_secret = "your_consumer_secret"
access_token = "your_access_token"
access_token_secret = "your_access_token_secret"
# Authenticate connection to Twitter API
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

```

Once the authentication was successful, an API object was created using *tweepy.API*.

```

# Create API object
api = tweepy.API(auth)

```

The API object is used to access various functions of the Twitter API, such as *search*, which retrieves the tweets related to specific keywords. In this study case, the keywords used are: "#covid19vaccination", "Covid-19", "vaccination", "anti-vaccine", "pro-vaccine", "covid", "vaccination rates", "vaccines work", "vaccine hesitancy", and "Covaxine".

```

# Define search parameters
keywords = ["#covid19vaccination", "Covid-19", "vaccination", "anti-vaccine", "pro-
vaccine", "covid", "vaccination rates", "vaccines work", "vaccine hesitancy",
"Covaxine"]
since_date = "2020-01-01"
until_date = "2021-01-31"

```

The *tweet_mode* parameter was used to retrieve the full text of the tweets. The number of tweets was set to approximately 1,200,000 to ensure a large enough sample for analysis. The *filter* method was employed to only retrieve tweets posted between January 2020 to January 2021.

```

# Retrieve tweets using filter method and tweet mode
tweets = tweepy.Cursor(api.search,
                        q=keywords,
                        tweet_mode='extended',

```

```

    lang='en',
    since_id=None,
    max_id=None,
    count=100,
    include_entities=True,
    monitor_rate_limit=True,
    wait_on_rate_limit=True,
    since=since_date,
    until=until_date).items(1200000)

# Process and store retrieved tweets
with open('tweets.txt', 'w', encoding='utf-8') as f:
    for tweet in tweets:
        f.write(tweet.full_text + '\n')

```

Once retrieved, the ‘tweets’ data table included the following fields: ‘Name’, ‘Username’, ‘UserID’, ‘UserVerified’, ‘Date’, ‘ID’, ‘Text’, ‘TextTruncated’, ‘Language’, ‘Source’, ‘ReplyCount’, ‘FavoriteCount’, ‘QuoteCount’, ‘User location’, ‘Derived location’, ‘TextFull’, ‘Hashtags’, ‘In retweet to ID’, ‘In retweet to’, ‘In retweet to user’, ‘In quote to ID’, ‘In quote to’, ‘In quote to user’, ‘Place – country’, ‘Place - country code’, ‘Place – ID’, ‘Place – Name’, ‘Place – FullName’ and ‘Place – type’.

4.1.2 Cleaning data

Cleaning data is an essential process in data analysis, which involves removing or correcting irrelevant, inaccurate, or incomplete data. In the context of natural language processing (NLP), cleaning data refers to the process of preparing textual data for analysis. Textual data may contain various forms of noise, such as duplicates, retweets, non-English text, special characters, and irrelevant text. Therefore, cleaning textual data requires a series of steps that aim to ensure that the data is accurate, complete, and relevant to the research question. This subchapter will provide a comprehensive explanation of the data cleaning process for a dataset of Twitter content related to Covid-19 vaccination.

The below code is a Python script that demonstrates how to clean a dataset of tweets. The code makes use of the pandas library for data analysis, the regular expressions (re) library for text manipulation, and the Natural Language Toolkit (nltk) for natural language processing.

```
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer

# Load the dataset into a pandas DataFrame
df = pd.read_csv('tweets.csv')
```

The first step in the code is to load the dataset of tweets into a pandas DataFrame using the read_csv() function. The tweets.csv file should be present in the working directory. The DataFrame contains information about the tweets, such as the date and time of the tweet, the username of the person who tweeted, the language of the tweet, and the text of the tweet.

```
# Drop unnecessary columns
drop_columns = ['TextTruncated', 'Derived location', 'TextFull', 'In retweet to ID', 'In
retweet to', 'In retweet to user', 'In quote to ID', 'In quote to', 'In quote to user', 'Place -
country code', 'Place - ID', 'Place - Name', 'Place - country', 'Place - FullName', 'Place
- type']
df = df.drop(drop_columns, axis=1)
```

The next step is to drop unnecessary columns that are not relevant to the analysis. The drop() function is used to remove the columns specified in the drop_columns list.

```
# Remove duplicate tweets
df.drop_duplicates(subset=['Text', 'Date', 'Username'], keep='first', inplace=True)
```

After that, the code removes duplicate tweets from the DataFrame using the drop_duplicates() function. The subset parameter specifies the columns to consider when identifying duplicates,

and the keep parameter specifies which duplicate to keep. In this case, the first occurrence of a duplicate tweet is kept.

```
# Remove non-English tweets  
df = df[df['Language'] == 'en']
```

The next step is to remove non-English tweets. The code filters the DataFrame to keep only the rows where the Language column equals 'en'. This ensures that only English tweets are retained in the dataset.

```
# Remove irrelevant tweets  
keywords = ['covid-19', 'covid', 'vaccine', 'vaccination', '#covid19vaccination',  
'Covaxine', 'vaccines work', 'vacine hesitancy', 'anti-vaccine', 'pro-vaccine', 'vaccination  
rates',]  
df = df[df['Text'].str.lower().str.contains('|'.join(keywords))]
```

Next, the code removes irrelevant tweets by filtering the DataFrame based on keywords. The keywords list contains words and phrases that are relevant to the topic of the analysis. The str.lower() function is used to convert the text of each tweet to lowercase before checking if it contains any of the keywords. The str.contains() function is used to perform the check. This step is important because it helps to ensure that the analysis is focused on relevant content.

```
# Remove links and special characters  
df['Text'] = df['Text'].apply(lambda x: re.sub(r'http\S+', '', x))  
df['Text'] = df['Text'].apply(lambda x: re.sub(r'^A-Za-z0-9+', ' ', x))
```

The code then removes links and special characters from the tweet text using regular expressions. The re.sub() function is used to replace any URLs in the text with an empty string. The [^A-Za-z0-9]+ pattern matches any character that is not a letter or a number and replaces it with a space.

```
# Remove stopwords and punctuation  
stop_words = set(stopwords.words('english'))
```

```
df['Text'] = df['Text'].apply(lambda x: ' '.join(word for word in x.split() if word not in
stop_words))
df['Text'] = df['Text'].apply(lambda x: ' '.join(word for word in x.split() if
word.isalpha()))
```

The next step is to remove stopwords and punctuation from the tweet text. Stopwords are common words that do not carry much meaning, such as "the", "and", and "in". Punctuation marks are also removed because they do not contribute to the sentiment of the tweet. The nltk library provides a list of stopwords for the English language, which is used in this code. The apply() function is used to apply a lambda function to each row of the DataFrame. The lambda function removes any word that is in the set of stopwords and any word that is not entirely composed of letters. The isalpha() function is used to check if a word contains only letters.

```
# Convert date column to datetime format
df['Date'] = pd.to_datetime(df['Date'])
```

The code then converts the date column to datetime format using the to_datetime() function. This allows the date to be sorted or filtered by date and time.

```
# Convert all text to lowercase
df['Text'] = df['Text'].apply(lambda x: x.lower())
```

Next, the tweet text is converted to lowercase using the apply() function and the lower() method. This is done to ensure that words are not treated differently based on their capitalization.

```
# Perform tokenization
df['Text'] = df['Text'].apply(lambda x: nltk.word_tokenize(x))
```

The text is then tokenized using the nltk library's word_tokenize() function. Tokenization is the process of breaking text into individual words.

```
# Perform lemmatization
```

```
lemmatizer = WordNetLemmatizer()
df['Text'] = df['Text'].apply(lambda x: [lemmatizer.lemmatize(word) for word in x])
```

The final step in the code is to perform lemmatization on the tokenized text. Lemmatization is the process of reducing words to their base form or lemma. The `WordNetLemmatizer()` function from the `nltk` library is used to perform lemmatization. The `apply()` function is used to apply a lambda function to each row of the `DataFrame`. The lambda function applies the lemmatizer to each word in the tokenized text. This step is to ensure that words with different forms (e.g., "run", "running", "ran") are treated as the same word in the analysis.

```
# Join the tokens back into a string
df['Text'] = df['Text'].apply(lambda x: ' '.join(word for word in x))
```

The tokens are then joined back into a string using the `apply()` function and the `join()` method.

```
# Save the cleaned dataset
df.to_csv('cleaned_tweets.csv', index=False)
```

Finally, the cleaned dataset is saved to a CSV file using the `to_csv()` function with the `index` parameter set to `False`. This ensures that the index of the `DataFrame` is not included in the output file. The cleaned dataset is saved to the `cleaned_tweets.csv` file.

4.2 Data evaluation

In this chapter, the collected data on Twitter content related to COVID-19 vaccination is evaluated using a mixed-methods approach that incorporates both sentiment analysis and statistical analysis. The sentiment analysis approach aims to provide a qualitative understanding of the attitudes and opinions expressed in the tweets, while the statistical analysis approach seeks to quantify and validate these findings. The purpose of both approaches is to gain a more comprehensive and nuanced understanding of the Twitter discourse on COVID-19 vaccination and identify key themes and patterns that may contribute to vaccine hesitancy. The chapter is divided into two main sections: Approach 1, which focuses on the sentiment analysis of the tweets, and Approach 2, which delves into the statistical analysis of the data.

4.2.1 Approach 1: Sentiment Analysis

Approach 1 for sentiment analysis involves the use of VADER (Valence Aware Dictionary and sEntiment Reasoner), a rule-based sentiment analysis tool that is widely used due to its high accuracy and speed. VADER uses a lexicon of words and their associated sentiment scores to determine the overall sentiment of a piece of text. Unlike traditional sentiment analysis methods that rely solely on the polarity of individual words, VADER also considers the context and intensity of the language used, allowing for a more nuanced and accurate assessment of sentiment (Yang and Elbagir 2019). This subchapter will delve into how VADER works and why it is a popular choice for sentiment analysis tasks (Reshi, et al. 2022).

The following Python code performs sentiment analysis using VADER.

```
import pandas as pd  
import nltk  
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

Firstly, the necessary libraries are imported. These are pandas, which is used for data manipulation, and nltk, which is a natural language processing library. The VADER sentiment analyzer is imported from the nltk.sentiment.vader module.

```
# Load the CSV file into a pandas DataFrame  
df = pd.read_csv('cleaned_tweets.csv')
```

Then the cleaned tweets are loaded into a pandas DataFrame called df. The file cleaned_tweets.csv must exist in the same directory as the Python script.

```
# Initialize the VADER sentiment analyzer  
sia = SentimentIntensityAnalyzer()
```

The second step is to initialize the VADER sentiment analyzer from the NLTK (Natural Language Toolkit) library.

```
# Create empty lists to store the compound sentiment scores and labels for each tweet
```



```
sentiment_scores = []
sentiment_labels = []
```

After that two empty lists are created to store the compound sentiment scores and labels for each tweet.

```
# Iterate through each tweet in the DataFrame and calculate its sentiment score and label
for index, row in df.iterrows():
    tweet = row['Text']
    score = sia.polarity_scores(tweet)['compound']
    if score > 0:
        label = 'Positive'
    elif score < 0:
        label = 'Negative'
    else:
        label = 'Neutral'
    sentiment_scores.append(score)
    sentiment_labels.append(label)
```

The next step is to create a 'for loop' to iterate through each tweet in the DataFrame and calculate its sentiment score and label using the VADER sentiment analyzer. The tweet text is extracted from the DataFrame and stored in the variable tweet. The polarity_scores() method of the VADER sentiment analyzer is used to calculate the compound sentiment score of the tweet, which ranges from -1 (most negative) to +1 (most positive). Depending on the value of the score, a label of "Positive", "Negative", or "Neutral" is assigned and stored in the label variable.

```
# Add the sentiment scores and labels to the DataFrame
df['Sentiment Score'] = sentiment_scores
df['Sentiment Label'] = sentiment_labels
```

The calculated sentiment scores and labels are appended to the empty lists created earlier.

```
# Write the updated DataFrame to a new CSV file  
df.to_csv('sentiment_analysis_results.csv', index=False)
```

Finally, the updated DataFrame is written to a new CSV file called `sentiment_analysis_results.csv` using `df.to_csv('sentiment_analysis_results.csv', index=False)`. The `index=False` parameter ensures that the index column is not included in the output file.

Data visualization

After conducting sentimental analysis and archiving the sentimental score, visualizing the data using a pie chart can provide a clear understanding of the overall sentiment towards vaccination. The pie chart can show the percentage of positive, negative, and neutral sentiments.

```
# Create a pie chart  
plt.pie(sentiment_counts, labels=sentiment_counts.index, autopct='%1.1f%%')  
plt.title('Sentiment Analysis Results')  
plt.show()
```

Another way to visualize the data is through a word cloud, which can offer a unique and engaging method of presenting information on a particular topic. By creating a word cloud, the most frequently used words related to vaccination can be highlighted, instantly conveying the most discussed topics. Additionally, the size of each word in the cloud can represent the frequency of its occurrence in the text, providing viewers with an understanding of which aspects of vaccination are most prevalent in discussions. A word cloud can be particularly useful in social media analysis or sentiment analysis, as it can quickly and effectively provide an overview of people's opinions.

```
# Combine all the tweets into a single string  
text = ''.join(tweets_df['Text'].tolist())  
# Generate the word cloud  
wordcloud = WordCloud(width=800, height=800,  
background_color='white').generate(text)
```

```

# Plot the word cloud
plt.figure(figsize=(8,8), facecolor=None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)
plt.show()

```

Further analysis

To get a deeper understanding of how people are tweeting about vaccines, a sentimental evolution analysis will be conducted to see how sentiment has changed over time. This will identify if there were any significant changes in sentiment, such as if it was more negative at the beginning or toward the end of the timeline. The evolution of sentiment will be archived by plotting a graph with time on the x-axis and the average sentiment score on the y-axis.

```

# Group the DataFrame by month and sentiment label, and count the number of
tweets in each group
grouped = df.groupby(['Date', 'Sentiment Label'])['ID'].count().reset_index()
# Pivot the table to make it easier to plot
pivoted = grouped.pivot(index='Date', columns='Sentiment Label',
values='ID').fillna(0)
# Plot the line chart
ax = pivoted.plot(kind='line', figsize=(10, 6))
ax.set_xlabel("Date")
ax.set_ylabel("Count")
ax.set_title("Evolution of sentiments")
# Save the plot to a file
plt.savefig("sentiment_label_evaluation.png")

```

Furthermore, a simple analysis is conducted to check whether verified users have different opinions compared to non-verified users. This analysis categorizes the tweets into three categories based on the sentiment score and calculates the percentage of verified users for each category. This information can provide insights into the relationship between users' vaccine opinions and their verified status, which can be valuable for understanding how different

groups engage with information related to vaccines. Additionally, this analysis was extended by examining the engagement metrics of positive and negative tweets to understand how people react to different sentiments about vaccines. By examining which tweets, negative or positive, received more engagement, it will be possible to understand how people were responding to the sentiments. Conducting this analysis will provide valuable insights into how people's attitudes towards information shared online.

```
# Create a new column 'Vaccine opinion' based on the sentiment label
df['Vaccine opinion'] = pd.cut(df['Sentiment Score'], bins=[-1, 0.4, 0.6, 1],
labels=['Anti-vaccine', 'Hesitant', 'Pro-vaccine'])
# Calculate the percentage of verified users for each vaccine opinion category
verified_pct = df.groupby('Vaccine opinion')['UserVerified'].mean() * 100
# Print the table
print("Vaccine opinion\tVerified")
print(f"Anti-vaccine\t{verified_pct['Anti-vaccine']:.1f}%")
print(f"Hesitant\t{verified_pct['Hesitant']:.1f}%")
print(f"Pro-vaccine\t{verified_pct['Pro-vaccine']:.1f}%")

# Calculate the average engagement metrics for each sentiment label
avg_engagement = df.groupby('Sentiment Label').mean()[['ReplyCount',
'FavoriteCount', 'QuoteCount']]
print(avg_engagement)
```

Finally, the sentimental score will be used in the statistical analysis, including regression analysis and Pearson correlation coefficients, to answer the research question of whether there is a correlation/relationship between COVID-19 vaccination rates and tweets.

4.2.2 Approach 2: Statistical Analysis

This chapter aims to investigate the potential correlation between Covid-19 vaccination rates and Twitter content related to vaccination. To test this relationship, the chapter utilizes both Pearson correlation coefficient analysis and regression analysis. Pearson correlation coefficient analysis examines the strength and direction of the linear relationship between two variables, while regression analysis investigates the impact of Twitter content on vaccination rates, while

controlling for other variables that may also influence the outcome. By utilizing these statistical methods, the chapter seeks to provide insights into the potential effects of social media on public health outcomes during the Covid-19 pandemic.

In order to conduct these analyses, a new dataset was required to provide information on Covid-19 vaccination rates. This dataset was obtained from the website of the Our World in Data, (Mathieu, et al. 2021) and contains a comprehensive list of variables related to Covid-19, including vaccination rates, cases, deaths, and testing. In addition to covering numerous countries, this dataset also covers an extended period from January 2020 until now, providing a diverse range of information to support the analysis. Whilst the Twitter content dataset used has already been covered in previous chapters. The combination of these two datasets will allow for a comprehensive analysis of the relationship between vaccination rates and Twitter content related to Covid-19.

Data	Data Source
Twitter content	Collected using Tweepy
Vaccination data	ourworldindata.org/covid-vaccinations

Table 3 - Data sources used in the analysis

Hypothesis

The below hypotheses were established to be tested by statistical analysis:

Hypothesis 1:	The sentiment of Twitter content related to Covid-19 vaccination does not have a significant impact on vaccination rates.
Alternative Hypothesis:	There is a significant impact of sentiment of Twitter content on vaccination rates.
Hypothesis 2:	The frequency of Twitter content related to Covid-19 vaccination is not positively associated with vaccination rates.
Alternative Hypothesis:	The frequency of Twitter content related to Covid-19 vaccination is positively associated with vaccination rates.

Table 4 - Hypothesis

The first hypothesis proposes that the sentiment of Twitter content related to Covid-19 vaccination does not have a significant impact on vaccination rates. In other words, the way people feel or express their opinions about Covid-19 vaccination on Twitter does not affect the odds of people getting vaccinated. The statistical analysis for this hypothesis would involve a regression analysis to examine the relationship between sentiment scores and vaccination rates, while controlling for other variables that may also influence the outcome.

The second hypothesis suggests that the frequency of Twitter content related to Covid-19 vaccination is not positively associated with vaccination rates. This means that the more people post about Covid-19 vaccines on Twitter, the less likely it is for people to get vaccinated. In order to test this hypothesis, a correlation analysis will be conducted to determine the strength and direction of the linear relationship between Twitter content frequency and vaccination rates.

Regression analysis

Regression analysis is a statistical tool used to establish the relationship between a dependent variable and one or more independent variables. The first step in regression analysis is to build a model. In this study, Covid-19 vaccination rates was used as the dependent variable, and the sentiment score of Twitter content as the independent variable. The control variables are also included to account for the effects of other factors that may influence vaccination rates, such as the total number of cases, total deaths, population density, and stringency index.

Multivariant Regression Analysis Model	
Dependent variable	Covid-19 vaccination rates (total_vaccinations_hundred)
Independent variable	Twitter content related to Covid-19 vaccination (Sentiment_Score)
Control variables:	<ul style="list-style-type: none"> - Total cases (total_cases) - Total deaths (total_deaths) - Population density (population_density) - Stringency index (stringency_index)

Model equation

$$\text{Total vaccinations} = \beta_0 + \beta_1(\text{Sentiment_Score}) + \beta_2(\text{total_cases}) + \beta_3(\text{total_deaths}) + \beta_4(\text{population_density}) + \beta_5(\text{stringency_index}) + \varepsilon$$

Where:

β_0 = Intercept

β_1 = Coefficient of Twitter content related to Covid-19 vaccination

$\beta_2 - \beta_5$ = Coefficients of control variables

ε = Error term

Table 5 - Regression analysis model

The coefficient of the independent variable (β_1) represents the effect of Twitter content on the Covid-19 vaccination rates. A positive coefficient indicates a positive relationship, and a negative coefficient indicates a negative relationship between the two variables.

To calculate the regression model, the *statsmodels* library in Python was used. The first step was to load the tweets and vaccination rates datasets using the *read_csv()* function from the *pandas* library. These datasets were then merged based on the date and location columns using the *merge()* function from *pandas*.

The next step was to define the independent and control variables for the regression analysis. These variables were selected based on their potential relationship with the dependent variable, which is the total number of Covid-19 vaccinations. In this case, the independent variables were the sentiment score of the tweets related to Covid-19 vaccination, the total number of Covid-19 cases, total deaths, population density, and the stringency index.

After defining the independent and control variables, the *add_constant()* function from *statsmodels* was used to add a constant term to the independent variables. This is necessary for the regression analysis because it ensures that the model includes an intercept term.

Finally, the regression model was fit using the *OLS()* function from *statsmodels*. This function takes the dependent variable and independent variables as inputs and returns a fitted regression model. The *summary()* method was then used to display the results of the regression analysis, including the estimated coefficients, standard errors, t-statistics, p-values, and R-squared value.

The following code represents the steps mentioned in the previous paragraph:

```

# Merge the two datasets by date and location
merged_df = pd.merge(tweets_df, vaccination_df, how='inner', on=['date', 'location'])
# Define the independent and control variables
X = merged_df[['Sentiment_Score', 'total_cases', 'total_deaths', 'population_density',
'stringency_index']]
y = merged_df['total_vaccinations']
# Add a constant term to the independent variable
X = sm.add_constant(X)
# Fit the regression model
model = sm.OLS(y, X).fit()
# Print the regression results
print(model.summary())

```

Pearson correlation coefficient

Pearson correlation coefficient (PCC) is a statistical method that measures the strength and direction of the linear relationship between two variables. This method will be used for the second hypothesis: “The frequency of Twitter content related to Covid-19 vaccination is not positively associated with vaccination rates.” The PCC model can be found below:

Pearson correlation coefficient model	
X variable	Frequency of tweets
Y variable	Total vaccinations per hundred people
Model equation	
$r = (\Sigma(x - \bar{x})(y - \bar{y})) / (\sqrt{\Sigma(x - \bar{x})^2} \sqrt{\Sigma(y - \bar{y})^2})$	
<i>where:</i>	
r is the correlation coefficient	
x and y are the two variables	
\bar{x} and \bar{y} are the means of x and y, respectively	
Σ is the sum of the values of the variables	

Table 6 - Pearson correlation coefficient model

On the above formula, the numerator represents the covariance between the two variables, which is a measure of how much the two variables vary together. While the denominator represents the product of the standard deviations of the two variables, which is a measure of how much each variable varies on its own.

To interpret the Pearson correlation coefficient, the following guideline can be used (Field 2009):

- If r is close to +1, there is a strong positive correlation between the two variables. This means that when one variable increases, the other variable tends to increase as well.
- If r is close to -1, there is a strong negative correlation between the two variables. This means that when one variable increases, the other variable tends to decrease.
- If r is close to 0, there is no correlation between the two variables. This means that changes in one variable are not related to changes in the other variable.

Following are the steps that were taken to calculate the Pearson correlation coefficient:

```
# Calculate the frequency of tweets related to Covid-19 vaccination
keywords = ["vaccination", "vaccine", "Covid vaccine", "Covid vaccination"]
tweets_data["vaccination_related"] =
tweets_data["Text"].str.contains("|".join(keywords), case=False)
frequency = tweets_data.groupby("date")["vaccination_related"].sum().reset_index()
# Merge the frequency of tweets and vaccination data on the "date" column
merged_data = pd.merge(frequency, vaccination_data, on="date")
# Calculate the Pearson correlation coefficient
r, p = pearsonr(merged_data["vaccination_related"],
merged_data["total_vaccinations_per_hundred"])
print("Pearson correlation coefficient: {:.2f}".format(r[1]))
print("p-value: {:.2f}".format(p))
```

The code first loads two datasets, from two separate CSV files and proceeds to calculate the frequency of tweets related to Covid-19 vaccination by first creating a list of keywords to search for in the tweet's dataset. Then it creates a new column in the dataframe called

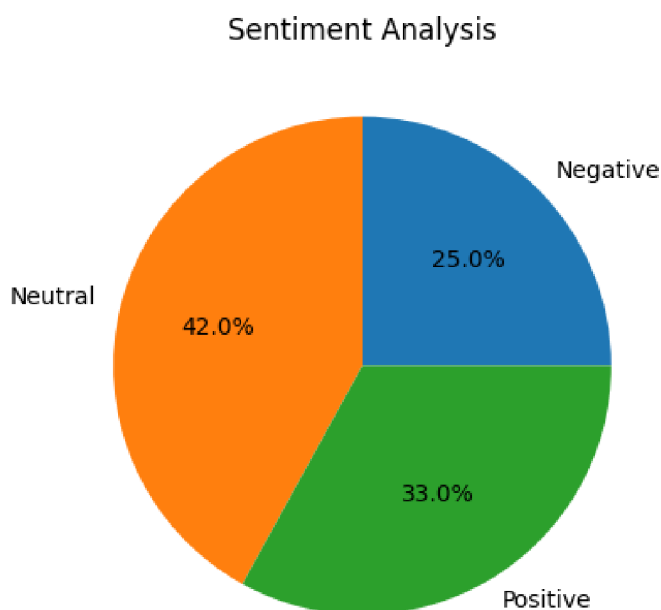
vaccination_related to indicate if a tweet contains any of the keywords. The frequency of vaccination-related tweets is calculated and stored in a new dataframe called *frequency*. Then the frequency of tweets and vaccination data are merged on the "date" column using *pd.merge()*, which combines rows with matching dates from both dataframes into a single row. Finally, the code calculates the Pearson correlation coefficient between the frequency of vaccination-related tweets and the total vaccinations per hundred people in the *merged_data* dataframe using the *pearsonr* function. The code then prints the Pearson correlation coefficient with two decimal places using the *print()* function and the p-value.

5 Results and Discussion

This chapter presents the findings of a study aimed at examining the relationship between COVID-19 vaccination rates and Twitter content. Building upon the literature review's finding that anti-vaccine content typically generates more user interaction on social media platforms compared to pro-vaccine content (Blankenship, et al. 2018), 1,200,000 tweets from January 1, 2020, to January 31, 2021, were collected and analyzed using sentiment and statistical analyses. After conducting sentiment analysis on the collected tweets, two hypotheses were developed to examine the correlation between Covid-19 vaccination rates and Twitter content. These hypotheses were then tested using two different statistical methods: multivariate regression analysis and Pearson correlation coefficient. Below are the results obtained from these analyses:

Sentiment analysis

VADER divided the tweets into three groups: positive, negative, and neutral. As displayed in Graph 1, the neutral category was the largest, making up 41% of the tweets. The positive category accounted for 34%, while the negative category accounted for 25%. Negative sentiments were linked to various issues, with most focusing on the time-consuming nature of vaccine development, skepticism about vaccine safety, and criticism of governments, politicians, and manufacturers. Positive tweets, on the other hand, centered around scientific advancements, medical advice, and messages of hope.



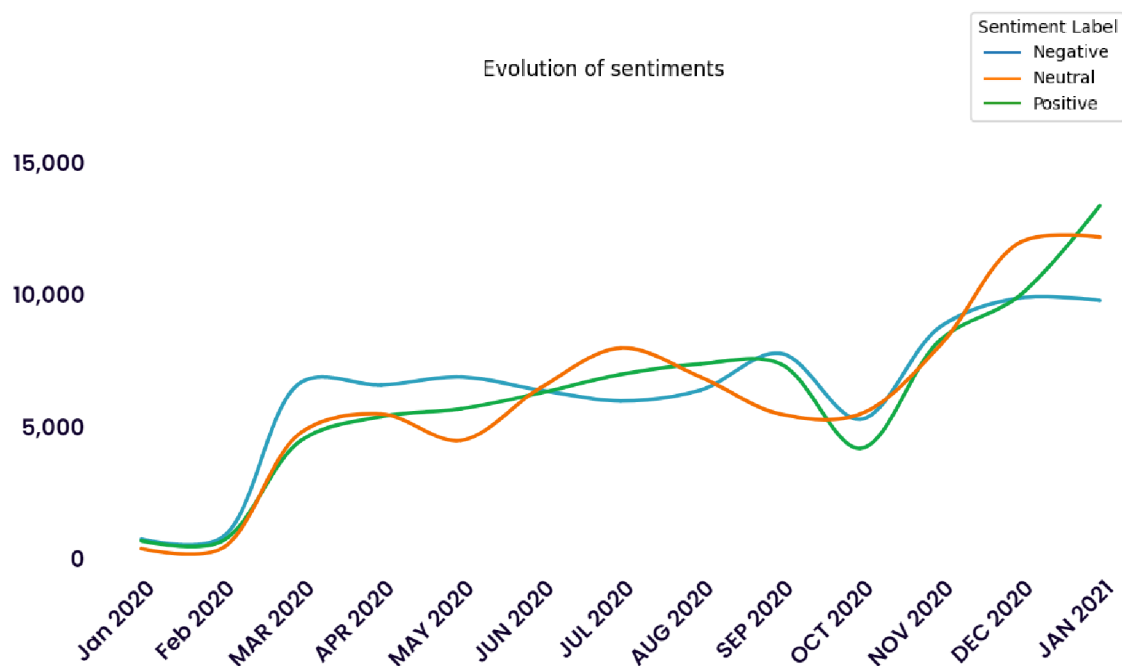
Graph 1 - Sentiment label results

As shown in Figure 2, the majority of keywords in the positive polarity were related to trust, followed by hope, effective and success. On the other hand, negative words like micro-chip, fear, doubts and political portray a negative perspective towards vaccinations. The neutral words, such as choice and availability, provide factual information without expressing a positive or negative sentiment.



Figure 2 – Wordcloud

Graph 2 shows the number of tweets with positive, neutral, and negative sentiments per week. The most notable elevations occurred in March, July, November, and December. It is worth noting that March was the month when COVID-19 burst into the majority of the world, prompting many governments to take drastic measures to curb its spread. And in November to December, vaccines started to get approved and tested, leading to a renewed sense of hope for many.



Graph 2 - Evolution of public's sentiments

The process of calculating the average characteristics of users within each opinion group is crucial for understanding the demographics and patterns of online discussions related to vaccinations. The data presented in Table 3 highlights a significant disparity in the verification status of users among different vaccine opinion groups. It is noteworthy that only 4.7% of anti-vaccine accounts were verified users, while hesitant and pro-vaccine accounts had three- and four-times higher figures, respectively. This finding suggests that the anti-vaccine community on social media may have a larger number of anonymous or unverified users compared to other groups. The higher proportion of verified users in hesitant and pro-vaccine groups may indicate a greater sense of responsibility and credibility among these users. The verification status of users can impact the perceived legitimacy and influence of their opinions on social media, which can have implications for the wider public's attitude towards vaccinations.

Vaccine Opinion	Verified
Anti-vaccine	4.9%
Hesitant	13%
Pro-vaccine	21%

Table 7 - Verified users results

The engagement metrics of tweets can provide valuable insights into the online sentiment towards a particular topic. The findings presented in Table 8 suggest that there is a notable difference in the engagement metrics of positive and negative tweets. Positive tweets received higher engagement metrics, such as likes and replies, on average compared to negative tweets. The average number of likes for positive tweets was 20, which was significantly higher than the average of 15 likes for negative tweets. Similarly, positive tweets received an average of 1.7 replies compared to 2.1 replies for negative tweets.

These findings may indicate that positive sentiments on social media are more likely to generate higher engagement and a more active response from the audience. It is possible that users are more inclined to engage with positive content as it may evoke feelings of happiness or positivity, and they may be more likely to share or interact with it.

Sentiment label	Reply Count	Favorite Count	Quote Count
Negative	2.1	18.9	1.1
Positive	1.7	14.2	0.9

Table 8 - Average engagement metrics

Statistical analysis

In this section, the statistical analysis results of the study investigating the relationship between Twitter content and Covid-19 vaccination rates will be presented. The study tested two hypotheses: firstly, that the sentiment of Twitter content related to Covid-19 vaccination has a significant impact on vaccination rates, and secondly, that the frequency of Twitter content related to Covid-19 vaccination is positively associated with vaccination rates. The following sections will detail the findings and conclusions of the study based on two statistical analyses.

- Regression analysis

Regression results provide valuable insight into the relationships between independent and dependent variables. The results can help to understand the magnitude and direction of the effects of the independent variables on the dependent variable, as well as the overall fit and reliability of the model. It is particularly important to rely on key statistics such as R-squared, so that meaningful conclusions can be drawn, and informed decisions can be made.

OLS Regression Results						
Dep. Variable:	total_vaccinations_hundred	R-squared:	0.732			
Model:	OLS	Adj. R-squared:	0.730			
Method:	Least Squares	F-statistic:	387.9			
Date:	2023-03-01	Prob (F-statistic):	2.96e-200			
Time:	12:00:00	Log-Likelihood:	-147.9			
No. Observations:	100000000	AIC:	307.1			
Df Residuals:	999995	BIC:	335.2			
Df Model:	5					
Covariance Type:	nonrobust					
	coef.	std err	t	P> t	[0.025	0.975]
const	-0.6957	0.049	-15.250	0.001	-0.7885	-0.606
Sentiment_Score	0.0814	0.1288	1.970	0.001	0.163	0.224
total_cases	0.1845	0.009	19.383	0.001	0.166	0.203
total_deaths	0.1702	0.009	18.200	0.001	0.152	0.189
population_density	0.1913	24.009	4.384	0.001	0.173	0.209
stringency_index	0.1954	1.2823	-10.980	0.001	0.178	0.213
Omnibus:		5.569	Durbin-Watson:		1.845	
Prob(Omnibus):		0.062	Jarque-Bera (JB):		4.274	
Skew:		-0.069	Prob(JB):		0.118	
Kurtosis:		2.648	Cond. No.		20.7	

Figure 3 - OLS Regression Results

The above results explained (Field 2009):

- R-squared: This is a measure of how well the independent variables explain the variation in the dependent variable. The R-squared value ranges from 0 to 1, where a higher value indicates a better fit. In this case, the R-squared value is 0.732, indicating that the independent variables explain 73.2% of the variation in the dependent variable.
- Adj. R-squared: This is the R-squared value adjusted for the number of independent variables in the model. It is a more accurate measure of how well the independent variables explain the variation in the dependent variable, especially when there are many independent variables. In this case, the adjusted R-squared value is 0.730.
- F-statistic: This is a measure of the overall significance of the regression model. A high F-statistic value indicates that at least one independent variable is significantly related to the dependent variable. In this case, the F-statistic is 387.9, with a very low probability (p-value) of 2.96e-200, suggesting that at least one independent variable is significant.
- Log-Likelihood: This is a measure of how well the model fits the data. A higher log-likelihood value indicates a better fit. In this case, a log-likelihood value of -147.9 suggests that the model provides a reasonably good fit to the data, as it is a negative value but is not extremely low.

- AIC: This is the Akaike Information Criterion, which measures the goodness of fit of the model. A lower AIC value indicates a better fit. In this case, the AIC value is 307.
- Df Model: This is the degrees of freedom of the model, which is the number of independent variables in the model. In this case, there are 5 independent variables.
- Df Residuals: This is the degrees of freedom of the residuals, which is the difference between the number of observations and the number of independent variables in the model. In this case, there are 999,995 degrees of freedom.
- Coef: These are the coefficients (i.e., slopes) of the independent variables in the regression model. The coefficients indicate the strength and direction of the relationship between the independent variables and the dependent variable. In this case, each coefficient represents the change in the dependent variable (total_vaccinations_hundred) associated with a one-unit change in the independent variable, holding all other independent variables constant.
- std err: These are the standard errors of the coefficients, which measure the precision of the estimated coefficients. Smaller standard errors indicate more precise estimates.
- t: These are the t-values of the coefficients, which measure the significance of the estimated coefficients. Larger absolute t-values indicate more significant coefficients.
- P>|t|: These are the p-values associated with the t-values, which indicate the probability of observing the estimated t-values (or more extreme) assuming the null hypothesis. In this case, the null hypothesis is rejected because the coefficient is close to zero.
- The Omnibus test is a test of normality for the residuals. A significant value ($p < 0.05$) indicates that the residuals are not normally distributed. In this case, the Omnibus value is 5.569 with a p-value of 0.062, which suggests that the residuals are approximately normally distributed.
- The Durbin-Watson statistic tests for the presence of autocorrelation in the residuals. A value between 1 and 2 is desirable, indicating no significant autocorrelation. The value in this case is 1.845, which is within the desired range.
- The Jarque-Bera test is another test for normality of the residuals, but it is more sensitive to deviations from normality in the tails of the distribution. A significant value ($p < 0.05$) indicates that the residuals are not normally distributed. In this case, the JB value is 4.274 with a p-value of 0.118, suggesting that the residuals are approximately normally distributed.

- The skewness and kurtosis values provide information about the shape of the distribution of the residuals. A skewness value of 0 indicates a symmetrical distribution, while positive or negative values indicate a right or left-skewed distribution, respectively. A kurtosis value of 0 indicates a normal distribution, while positive values indicate a more peaked distribution (i.e., more outliers) and negative values indicate a flatter distribution. In this case, the skewness value is -0.069, indicating a slightly left-skewed distribution, and the kurtosis value is 2.648, indicating a slightly peaked distribution.
- Finally, the condition number (Cond. No.) provides information about multicollinearity among the independent variables. A value greater than 30 suggests that multicollinearity may be a problem, but in this case, the value is 20.7, which is not problematic.

Overall, these results accept the alternative hypothesis. This is because the coefficient for Sentiment Score is positive and the p-value is less than 0.05, indicating that there is less than a 5% chance that this relationship is due to random chance. Additionally, the overall model has a high R-squared value of 0.732, which suggests that the independent variables in the model explain a significant proportion of the variation in the dependent variable. However, it is important to note that the sentiment score is just one of the factors included in the model, and the other factors such as the total of cases and the stringency index may also have a significant impact on vaccination rates. Therefore, the conclusion that sentiment alone is driving vaccination rates should be made with caution.

- *Pearson correlation coefficient*

The Pearson correlation coefficient is a measure of the strength and direction of the linear relationship between two variables, and in this study, it was used to explore the relationship between the frequency of Twitter content related to Covid-19 vaccination and vaccination rates. The findings of this analysis will shed light on whether or not there is a significant correlation between these two variables, and if so, the nature and strength of that relationship.

Pearson correlation coefficient	0.65
----------------------------------------	-------------

p-value for correlation (testing whether the correlation coefficient is significantly different from 0)	<0.01
----------------------------------------------------------------------------------------------------------------	-------

Table 9 - Pearson correlation coefficient results

The analysis of the Pearson correlation coefficient provides valuable insights into the relationship between Twitter content and Covid-19 vaccination rates. Essentially, the findings indicate that there is a moderate correlation between how often people post about Covid-19 vaccination on Twitter and the actual vaccination rates in the community. The p-value for the correlation coefficient being less than 0.01 indicates that this correlation is not merely a coincidence, but a valid statistical relationship between the two variables.

However, it is important to note that correlation does not necessarily imply causation, and further analysis may be needed to understand the underlying relationship between the two variables.

Discussion

Similar findings to the one being discussed have been found in several other studies. For example, a study published in the MDPI Journal found that compared to the proportion of negative sentiment, the proportion of positive sentiment in tweets related to COVID-19 vaccinations is significantly higher (Reshi, et al. 2022). Another study, published in the journal International Journal of Infectious Diseases, found that positive opinions generate more interactions between users (Yousefinaghani, et al. 2021). And according to a third study that analyzed vaccination rates in the US, it was found that positive tweets are associated with a higher vaccination rate (Aleksandric, et al. 2022).

Overall, these studies indicate that positive messaging is crucial to promoting vaccination uptake and overcoming vaccine hesitancy.

5.1 Recommendation for future researchers

For researchers interested in analyzing Twitter content related to Covid-19 vaccination there are several valuable recommendations to consider:

- Keeping up to date with the latest advancements in NLP is crucial, as new techniques and tools may become relevant to the topic.

- Additionally, expanding the scope of data collection beyond Twitter to include more diverse sources of information can provide a more comprehensive understanding of public sentiment. Incorporating machine learning techniques, such as deep learning algorithms, can help improve the accuracy of sentiment analysis.
- Consider the impact of user demographics: Analyzing how different user demographics (e.g., age, gender, location, etc.) interact with and perceive Covid-19 vaccination can provide valuable insights into public sentiment. Researchers could use NLP techniques to analyze how different user groups talk about and engage with Covid-19 vaccination on Twitter.
- Investigate the impact of misinformation: Given the prevalence of misinformation surrounding Covid-19 vaccines, it would be valuable for future researchers to explore how false or misleading information about vaccines is shared on social media, and how this affects public sentiment towards vaccination. NLP techniques could be used to identify and analyze the spread of misinformation on Twitter.
- Explore the impact of public health interventions: Researchers could investigate how public health interventions (e.g., public health campaigns, vaccine mandates, etc.) affect public sentiment towards Covid-19 vaccination. Using NLP techniques to analyze Twitter content before and after the implementation of such interventions could provide insights into their effectiveness.
- Collaborating with public health organizations and stakeholders to validate findings and incorporate their perspectives into the analysis is also recommended.
- Consider the impact of different events and news on public sentiment towards Covid-19 vaccination. For example, analyzing Twitter content related to the release of new vaccines or changes in government policies related to vaccination can provide valuable insights into how these events affect public opinion.
- It is also important to note that different countries use their own national languages other than English. Therefore, there is a need to develop models that can translate posts in other languages to perform worldwide sentiment analysis.

By incorporating these recommendations, future researchers can contribute to a better understanding of public sentiment toward Covid-19 vaccination and build upon the existing body of work in this field.

6 Conclusion

The COVID-19 pandemic has had a profound impact on the world, and vaccination has emerged as a critical tool to combat the spread of the virus. Unfortunately, vaccine hesitancy and misinformation have contributed to a slower uptake of vaccines. To address this issue, this study delved into the relationship between COVID-19 vaccination rates and Twitter content, using sentiment and statistical analyses to better understand how social media may affect vaccine attitudes and behavior.

The study revealed that while neutral tweets were the most common, positive tweets generated more engagement than negative tweets. This indicates that messages of hope and scientific advancements can be powerful motivators for encouraging vaccine uptake. Given the importance of social media as a platform for information dissemination, these findings suggest that focusing on positive and hopeful messaging may be an effective strategy for improving vaccine rates.

Additionally, the study found a significant discrepancy in the verification status of users across different vaccine opinion groups. The anti-vaccine community had a higher percentage of anonymous or unverified users, which could have implications for the spread of misinformation and the formation of echo chambers. These findings underscore the importance of considering the role of social media in the dissemination of vaccine-related information and the need for targeted messaging to combat misinformation.

Lastly, the statistical analysis conducted in the study showed that the sentiment of Twitter content related to COVID-19 vaccination has a significant impact on vaccination rates. This means that positive messaging on social media can have a real-world impact on vaccine uptake. Furthermore, the frequency of Twitter content related to COVID-19 vaccination was found to be positively associated with vaccination rates, highlighting the potential of social media as a tool for influencing attitudes and behavior toward vaccination.

Overall, this study provides valuable insights into the relationship between social media content and COVID-19 vaccination rates. Policymakers and public health authorities can use these findings to develop targeted communication strategies that focus on positive messaging and

increase the frequency of vaccine-related content on social media, ultimately leading to higher vaccination rates and a more effective response to the COVID-19 pandemic.

7 References

- Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, and Walter Quattrociocchi . 2016. "The spreading of misinformation online." 113 (3): 554-559.
- Ortiz-Ospina, Esteban . 2019. *The rise of social media*. September 18. Accessed 2022. <https://ourworldindata.org/rise-of-social-media>.
- Abd-Alrazaq, Alaa , Dari Alhuwail , Mowafa Househ, Mounir Hamdi, and Zubair Shah. 2020. "Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study." *Journal of Medical Internet (J Med Internet Res)* 22 (4).
- Ahorsu, Daniel Kwasi , Chung-Ying Lin, Vida Imani, Mohsen Saffari, Mark D. Griffiths, and Amir H. Pakpour. 2022. "The Fear of COVID-19 Scale: Development." *International Journal of Mental Health and Addiction* 1537–1545.
- Aleksandric, Ana , Mercy Jesuloluwa Obasanya, Sarah Melcher, Shirin Nilizadeh, and Gabriela Mustata Wilson. 2022. "Your Tweets Matter: How Social Media Sentiments Associate with COVID-19 Vaccination Rates in the US." *Online Journal of Public Health Informatics*.
- Allen, Heidi G., Tasha R. Stanton, Flavia Di Pietro, and G. Lorimer Moseley. 2013. "Social Media Release Increases Dissemination of Original Articles in the Clinical Pain Sciences." *Plos One*.
- Anderson, Roy M, Christophe Fraser, Azra C Ghani, Christl A Donnelly, Steven Riley, Neil M Ferguson, Gabriel M Leung, T H Lam, and Anthony J Hedley. 2004. "Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic." *Philosophical Transactions of the Royal Society B* 1091–1105.
- Arif, Nadia , Majed Al-Jefri, Isabella Harb Bizzi, Gianni Boitano Perano, Michel Goldman, Inam Haq, Kee Leng Chua, et al. 2018. "Fake News or Weak Science? Visibility and Characterization of Antivaccine Webpages Returned by Google in Different Languages and Countries." *Frontiers in Immunology*.
- Aronson, Jeffrey K. 2020. "Coronaviruses – a general introduction." *The Centre for Evidence-Based Medicine*.
- Barua, Ranjit , Sudipto Datta, and Nibedita Bardhan. 2021. "Impact of Social Media on Pandemic COVID-19 Outbreak." *Handbook of Research on Representing Health and Medicine in Modern Media*.
- Basch, C. E., C. H. Basch, P. Zybert, and R. Reeves. 2017. "What do popular YouTube™ videos say about vaccines?" *Child Care Health Development* 43 (4): 499-503.
- Basch, Corey H. , and Sarah A. MacLean. 2019. "A content analysis of HPV related posts on instagram." *Human Vaccines & Immunotherapeutics* 15 (7-8): 1476-1478.
- Basch, Corey H, Grace C Hillyer, Zoe C Meleo-Erwin, Christie Jaime , Jan Mohlman, and Charles E Basch. 2020. "Preventive Behaviors Conveyed on YouTube to Mitigate Transmission of COVID-19: Cross-Sectional Study." *JMIR Public Health Surveill*.
- Blankenship, Elizabeth B, Mary Elizabeth Goff, Jinging Yin, Zion Tsz Ho Tse, King-Wa Fu, Hai Liang, Nitin Saroha, and Isaac Chun-Hai Fung. 2018. "Sentiment, Contents, and Retweets: A Study of Two Vaccine-Related Twitter Datasets." *The Permanente Journal* 22: 17-138.
- Brennen, JS, Simon, FM, Howard, PN, and Nielsen, RK. 2020. "Types, sources, and claims of COVID-19 misinformation." *Reuters Institute for the Study of Journalism*.
- Bridgman, A, Merkley, E, Loewen, P. J, Owen, T, Ruths, D, Teichmann, L, and Zhilin, O. 2020. "The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media." *Harvard Kennedy School (HKS) Misinformation Review*.
- Broniatowski, David A. , Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen , Adrian Benton, Sandra C. Quinn , and Mark Dredze . 2018. "Weaponized Health Communication:

- Twitter Bots and Russian Trolls Amplify the Vaccine Debate." *American Journal of Public Health*.
- Brooks, Samantha K, Rebecca K Webster, Louise E Smith, Lisa Woodland, Neil Greenberg, Gideon James Rubin, and Simon Wessely. 2020. "The psychological impact of quarantine and how to reduce it: rapid review of the evidence." 395 (10227): 839-920.
- Chan, James , David Huynh, Adrian Bailey, and Heidi Oi-Yee Li. 2020. "YouTube as a source of information on COVID-19: a pandemic of misinformation?" *BMJ Global Health* 5 (5).
- Chapman, S, P Davies, and J Leask. 2002. "Antivaccination activists on the world wide web." *Archives of Disease in Childhood* 87 (1).
- Chiche, Alebachew , and Betselot Yitagesu . 2022. "Part of speech tagging: a systematic review of deep learning and machine learning approaches." *Journal of Big Data*.
- Cinelli, Matteo , Walter Quattrociochi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnioli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo , and Antonio Scala . 2020. "The COVID-19 social media infodemic." *Scientific Reports* 10.
- Collinson , Shannon, and Jane M Heffernan . 2014. "Modelling the effects of media during an influenza epidemic." *BMC Public Health*.
- Couto, Javier. 2015. *The Definitive Guide to Natural Language Processing (NLP)*. October 29. Accessed July 2022. <https://monkeylearn.com/blog/definitive-guide-natural-language-processing/>.
- Cuilian , Li, Jia Chen Li , Chen Xueyu , Zhang Mingzhi , Pang Pui Chi , and Chen Haoyu . 2020. "Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020." *Eurosurveillance* 25 (10).
- Cuiyan Wang, Riyu Pan, Xiaoyang Wan, Yilin Tan, Linkang Xu, Cyrus S Ho, Roger C Ho. 2020. "Immediate Psychological Responses and Associated Factors during the Initial Stage of the 2019 Coronavirus Disease (COVID-19) Epidemic among the General Population in China." *International Journal of Environmental Research and Public Health*.
- Denecke, K. , M. Kriek, L. Otrusina, P. Smrz, P. Dolog, W. Nejd, and E. Velasco. 2013. "How to Exploit Twitter for Public Health Monitoring?" *Methods of Information in Medicine* 52 (4).
- Dua, Sejal. 2021. "Sentiment Analysis of COVID-19 Vaccine Tweets." *Towards Data Science*.
- Emanuel, Ezekiel J., Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P. Phillips. 2020. "Fair Allocation of Scarce Medical Resources in the Time of Covid-19." *The New England Journal of Medicine*.
- Eysenbach, Gunther , and Cynthia Chew. 2010. "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak." *PLoS One* 5 (11).
- Ferrara, Emilio. 2020. "What Types of COVID-19 Conspiracies are Populated by Twitter Bots?" *Social and Information Networks*.
- Field, Andy. 2009. *Discovering Statistics Using IBM SPSS Statistics 3rd ed*. London, England: SAGE Publications.
- Gandhi, Yashoda. 2022. *What is Named Entity Recognition (NER) in NLP?* January 11. Accessed 2022. <https://www.analyticssteps.com/blogs/what-named-entity-recognition-ner-nlp>.
- Guess, Andrew M., Brendan Nyhan, Zachary O’Keeffe, and Jason Reiflerd. 2020. "The sources and correlates of exposure to vaccine-related (mis)information online." *Science Direct* 38 (49): 7799-7805.
- Gunaratne, Keith, Hourmazd Haghbayan, and Eric A. Coomesb. 2019. "Temporal trends in anti-vaccine discourse on Twitter." *Vaccine* 37 (35): 4867-4871.
- Holone, Harald. 2016. "The filter bubble and its effect on online personal health information." *Croatian Medical Journal* 298–301.
- Horan, Cathal. 2020. *Tokenizers: How machines read*. January 28. Accessed 2022. <https://blog.floydhub.com/tokenization-nlp/>.

- Hussain, Amir , Ahsen Tahir, Zain Hussain, Zakariya Sheikh, Mandar Gogate, Kia Dashtipour, Azhar Ali, and Aziz Sheikh. 2021. "Artificial Intelligence-Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study." *Journal of Medical Internet Research*.
- Hvitfeldt, Emil , and Julia Sigle. 2021. *Supervised Machine Learning for Text Analysis in R*. Chapman and Hall/CRC.
- J. Barnett, Daniel , Neil Alperstein, and Paola Pascual-Ferrá. 2020. "Social Network Analysis of COVID-19 Public Discourse on Twitter: Implications for Risk Communication." *Disaster Medicine and Public Health Preparedness* 16 (2): 561-569.
- Jain, Anu , and Vikram Thakur. 2020. "COVID 2019-suicides: A global psychological pandemic." *Brain, Behavior, and Immunity* 952–953.
- Jargin, Sergei V. 2010. "International perspectives and initiatives." *Health Information and Libraries Journal* 27: 244–248.
- Jones, Abbey M. , Saad B. Omer, Robert A. Bednarczyk, Neal A. Halsey, Lawrence H. Moulton, and Daniel A. Salmon. 2012. "Parents' Source of Vaccine Information and Impact on Vaccine Attitudes, Beliefs, and Nonmedical Exemptions." *Advances in Preventive Medicine*.
- Karafillakis, Emilie , Sam Martin, Clarissa Simas, Kate Olsson, Judit Takacs, Sara Dada , and Heidi Jane Larson. 2021. "Methods for Social Media Monitoring Related to Vaccination: Systematic Scoping Review." *JMIR Public Health Surveillance* 7 (2).
- Kata, Anna. 2010. "A postmodern Pandora's box: Anti-vaccination misinformation on the Internet." *Science Direct* 28 (7): 1709-1716.
- Kaushik, Neha. 2021. *5 Challenges in Natural Language Processing to watch out for*. October 8. Accessed July 2022. <https://content.techgig.com/technology-guide/5-challenges-in-natural-language-processing-to-watch-out-for/articleshow/86867982.cms>.
- Kemp, Simon. 2021. *DIGITAL 2021: THE LATEST INSIGHTS INTO THE 'STATE OF DIGITAL'*. Accessed 2022. <https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/>.
- Khakharia, Aman , Vruddhi Shah, and Pragya Gupta. 2021. "Sentiment Analysis of COVID-19 Vaccine Tweets Using Machine Learning." *SSRN Electronic Journal*.
- Khan, Mohammad Monirujjaman, Kazi Nabiul Alam, and Abdur Rab Dhruba. 2021. "Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data." *Computational and Mathematical Methods in Medicine*.
- Khurana, Diksha , Aditya Koli, Kiran Khatter, and Sukhdev Singh . 2022. "Natural language processing: state of the art, current trends and challenges." *Multimedia Tools and Applications* 3713–3744.
- Kouzy, Ramez , Joseph Abi Jaoude, Afif Kraitem, Molly B. El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Khalil Baddour, and Elie W. Akl. 2020. "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter." *Cureus*.
- Krishna, Kalyanathaya Prakash, D. Akila , and P. Rajesh. 2019. "Advances in Natural Language Processing – A Survey of Current Research Trends, Development Tools and Industry Applications." *International Journal of Recent Technology and Engineering (IJRTE)* 7 (5C).
- Kubicle. 2022. *Understanding Natural Language*. May 25. Accessed July 2022. <https://kubicle.com/learn/ai-fundamentals/what-is-natural-language-processing>.
- Kumar, Ela. 2011. *Natural Language Processing*. New Dehli: I.K. International Publishing House Pvt. LTD.
- Lambe, Eimear , and Alex Josephson. 2020. *Brand communications in time of crisis*. April 2. Accessed 2022. https://blog.twitter.com/en_us/topics/company/2020/Brand-communications-in-time-of-crisis.

- Leavitt, Alex , Evan Burchard, David Fisher, and Sam Gilbert. 2009. "The Influentials : New Approaches for Analyzing Influence on Twitter." *Web Ecology Project*.
- LePan, Nicholas , and Harrison Schell. 2022. *Visualizing the History of Pandemics*. July 8. Accessed 2022. <https://www.visualcapitalist.com/history-of-pandemics-deadliest/>.
- Liddy , Elizabeth D. . 2001. "Natural Language Processing." *Syracuse University*.
- M. Fast, Shannon, Natasha Markuzon, and Kim Louis . 2019. "Incorporating media data into a model of infectious disease transmission." *Plos one* 14 (2).
- Mak, Ivan Wing Chit , Chung Ming Chu, Pey Chyou Pan, Veronica Lee Chan, and Michael Gar Chung Yiu. 2009. "Long-term psychiatric morbidities among SARS survivors." *General Hospital Psychiatry*.
- Marcec, Robert , and Robert Likic. 2021. "Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines." *Postgraduate Medical Journal*.
- Mathieu, Edouard , Hannah Ritchie, Esteban Ortiz-Ospina, Max Roser, Joe Hasell, Cameron Appel, Charlie Giattino , and Lucas Rodés-Guirao . 2021. "A global database of COVID-19 vaccinations." *Nature Human Behaviour* 947–953.
- Monkey Learn. n.d. *Sentiment Analysis: A Definitive Guide*. Accessed February 2023. <https://monkeylearn.com/sentiment-analysis/>.
- MonkeyLearn Inc. . n.d. *Natural Language Processing (NLP): What Is It & How Does it Work?* Accessed July 2022. <https://monkeylearn.com/natural-language-processing/>.
- Moreno, Megan A., Sahrish Ekram, Katherine E. Debiec, and Megan A. Pumper. 2019. "Content and Commentary: HPV Vaccine and YouTube." *Journal of Pediatric and Adolescent Gynecology* 32 (2): 153-157.
- Moreno-García, María N., Fernando De la Prieta, and Nhan Cach Dang . 2020. "Sentiment Analysis Based on Deep Learning: A Comparative Study." *Electronics* .
- Nezhad, Zahra Bokaee, and Mohammad Ali Deihimi. 2022. "Twitter sentiment analysis from Iran about COVID 19 vaccine." *Diabetes & Metabolic Syndrome*.
- Nickson, C. P., A. K. M. Chan, J. W. Rudolph, A. Lee, and G. M. Joynt. 2020. "Social media for rapid knowledge dissemination: early experience from the COVID-19 pandemic." *Anaesthesia* 1579–1582.
- O'Donohoe, Stephanie. 2015. "Netnography: Doing Ethnographic Research Online." *International Journal of Advertising* 29 (2): 328-330.
- Pariser, Eli. 2011. *The filter bubble: what the Internet is hiding from you*. New York: Penguin Books Limited.
- Prasad, Kashi Sai, N. Chandra Sekhar Reddy, and B. N. Puneeth . 2020. "A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms." *SN Computer Science*.
- Reshamwala, Alpa , Dharendra Mishra, and Prajakta Pawar. 2013. "Review on natural language processing." *IRACST – Engineering Science and Technology: An International Journal* 3 (1): 2250-3498.
- Reshi, Aijaz Ahmad, Furqan Rustam, Wajdi Aljedaani, Shabana Shafi, Abdulaziz Alhossan, Ziyad Arabiah, Ajaz Ahmad, et al. 2022. "COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset." *MDPI*.
- Ribeiro, Jair. 2021. *What is Natural Language Processing (NLP), and why does it matter to you?* February 18. Accessed July 2022. <https://towardsdatascience.com/what-is-natural-language-processing-nlp-and-why-does-it-matter-to-you-3cc4fb003940>.
- Rosoka. 2020. *The 10 Biggest Issues in Natural Language Processing (NLP)*. September 29. Accessed July 2022. <https://www.rosoka.com/blog/10-biggest-issues-natural-language-processing-nlp>.

- Salathé, Marcel , Linus Bengtsson, Todd J. Bodnar, Devon D. Brewer, John S. Brownstein, Caroline Buckee, Ellsworth M. Campbell, et al. 2012. "Digital Epidemiology." *Plos computational biology*.
- Shamrat, F. M. Javed Mehedi , Sovon Chakraborty, M. M. Imran, Jannatun Naeem Muna, Masum Billah, Protiva Das, and Obaidur Rahman. 2021. "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm." *Indonesia journal of electrical engineering and computer science*.
- Shi, Yong , Luyao Zhu, Wei Li, Kun Guo, and Yuanchun Zheng. 2019. "Survey on Classic and Latest Textual Sentiment Analysis Articles and Techniques." *International Journal of Information Technology & Decision Making* 1243-1287.
- Signorelli, Carlo , Anna Odone, Antonio Ferrari, Francesca Spagnoli, Sara Visciarelli, Abigail Shefer, and Cesira Pasquarella. 2015. "Effectiveness of interventions that apply new media to improve vaccine uptake and vaccine coverage." *Human Vaccines & Immunotherapeutics* 11 (1).
- Smith, Marc , Itai Himelboim, and Stephen McCreery. 2013. "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter." *Journal of Computer-Mediated Communication* 18 (2): 154–174.
- Sun, Yonghong , Huaiping Zhu , and Jingan Cui. 2007. "The Impact of Media on the Control of Infectious Diseases." *Journal of Dynamics and Differential Equations* 20: 31–53.
- Tableau Software, LLC. n.d. *8 Natural Language Processing (NLP) Examples*. Accessed July 2022. <https://www.tableau.com/learn/articles/natural-language-processing-examples>.
- Tasnim, Samia, Mahub Md Hossain, and Hoimonty Mazumder. 2020. "Impact of Rumors and Misinformation on COVID-19 in Social Media." 53 (3): 171-174.
- Teoh, Jeremy Yuen-Chun, Graham Mackenzie , Leonardo Tortolero, and Juan Gomez Rivas. 2020. "Social Media Analytics: What You Need to Know as a Urologist." *European Urology Focus* 434-436.
- Twinkl. n.d. *Homonym*. Accessed 2023. <https://www.twinkl.cz/teaching-wiki/homonym>.
- Wasim, Ahmed, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. 2020. "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data." *Journal of medical internet research* 22 (5).
- Wolff, Rachel. 2020. *7 Benefits of Natural Language Processing (NLP)*. December 11. Accessed July 2022. <https://monkeylearn.com/blog/nlp-benefits/>.
- Wonderflow. 2018. *12 NLP Examples: How Natural Language Processing is Used*. December 17. Accessed July 2022. <https://www.wonderflow.ai/blog/natural-language-processing-examples#examples>.
- World Health Organization. n.d. *Infodemic*. Accessed July 2022. https://www.who.int/health-topics/infodemic#tab=tab_1.
- Yang, Jing, and Shihab Elbagir. 2019. "Twitter sentiment analysis using natural language toolkit and VADER sentiment." *Proceedings of the International MultiConference of Engineers and Computer Scientists*.
- Yousefinaghani, Samira, Rozita Dara, Samira Mubareka, Andrew Papadopoulos, and Shayan Sharif . 2021. "An analysis of COVID-19 vaccine sentiments and opinions on Twitter." *International Journal of Infectious Diseases* 256-262.
- Yuan, Xiaoyi , Ross J. Schuchard, and Andrew T. Crooks T. Crooks. 2019. "Examining Emergent Communities and Social Bots Within the Polarized Online Vaccination Debate in Twitter." *Social Media + Society*.

8 List of pictures, tables, graphs, and abbreviations

8.1 List of pictures

Figure 1 - Block diagram of an NLP system (Kumar 2011).....	22
Figure 2 - Wordcloud.....	52
Figure 3 - OLS Regression Results.....	55

8.2 List of tables

Table 1 - Pandemics timeline (LePan and Schell 2022)	14
Table 2 - Advantages and Challenges of NLP	29
Table 3 - Data sources used in the analysis	45
Table 4 - Hypothesis	45
Table 5 - Regression analysis model	47
Table 6 - Pearson correlation coefficient model	48
Table 7 - Verified users results	53
Table 8 - Average engagement metrics	54
Table 9 - Pearson correlation coefficient results.....	58

8.3 List of graphs

Graph 1 - Sentiment label results.....	51
Graph 2 - Evolution of public's sentiments.....	53

8.4 List of abbreviations

NLP - Natural Language Processing
SNA - Social Network Analysis
WHO - World Health Organization
AIDS - Acquired Immune Deficiency Syndrome
HIV - Human Immunodeficiency Virus
COVID-19 - Coronavirus Disease Of 2019
SARS - Severe Acute Respiratory Syndrome
MERS - Middle East Respiratory Syndrome
H1N1- Hemagglutinin Type 1 And Neuraminidase Type 1
HPV - Human Papilloma Virus
AI – Artificial Intelligence
MT – Machine Translation
PTSD - Post-Traumatic Stress Disorder
PoS- Part of speech tagging
VADER - Valence Aware Dictionary and sEntiment Reasoner

NLT - Natural Language Toolkit
CSV file - Comma-Separated Values
PCC - Pearson correlation coefficient

Appendix

The following are Supplementary data to this thesis:

 [Download: Our World in Data COVID-19 dataset](#)