UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE GRADO

STUDY AND DETECTION OF ANTI-VACCINE LANGUAGE ON SOCIAL MEDIA USING NATURAL LANGUAGE PROCESSING TECHNIQUES

Mª FELIPA LEDESMA CORNIEL

JUNIO 2023

TRABAJO DE FIN DE GRADO

Título:	Estudio y Detección de Lenguaje Anti-Vacunas en Redes Sociales usando Técnicas de Procesamiento del Lenguaje Natural
Título (inglés):	Study and Detection of Anti-Vaccine Language on Social Media using Natural Language Processing Techniques
Autor:	Mª Felipa Ledesma Corniel
Tutor:	Óscar Araque Iborra
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

STUDY AND DETECTION OF ANTI-VACCINE LANGUAGE ON SOCIAL MEDIA USING NATURAL LANGUAGE PROCESSING TECHNIQUES

Mª Felipa Ledesma Corniel

Junio 2023

Resumen

Este proyecto presenta un análisis del discurso anti-vacunas en la red social Twitter mediante el uso de técnicas de Procesamiento del Lenguaje Natural (NLP). Las vacunas son tecnologías biomédicas que estimulan el sistema inmunológico para generar una respuesta protectora contra enfermedades específicas. Sin embargo, a pesar de haber demostrado su eficacia a lo largo de la historia, generan desconfianza y temor en una parte de la población.

Para hacer un estudio de este movimiento extraemos mensajes anti-vacunas de Twitter. Las redes sociales ofrecen una valiosa fuente de información sobre la que realizar técnica de NLP, debido a la abundancia de datos y a la diversidad de perspectivas que se comparten en ellas en tiempo real. Se lleva a cabo una investigación exploratoria para identificar y comprender cómo se comunican estas comunidades en Twitter. A continuación, se selecciona y recopila cuidadosamente un conjunto de datos compuesto por tweets de los últimos años para su análisis posterior.

En primer lugar, se realiza un preprocesamiento del texto para garantizar la calidad y la integridad del conjunto de datos. Seguidamente, se emplean técnicas avanzadas de NLP para descurbir patrones y estructuras lingüísticas en los tweets recopilados. Estas técnicas incluyen la extracción de n-gramas y el análisis temporal de vocabulario para identificar los cambios en el uso del lenguaje a lo largo del tiempo. Además, se realiza un análisis comparativo para obtener las diferencias en perspectivas y argumentos entre anti-vacunas y pro-vacunas. Para profundizar más, se utilizan modelos de lenguajes preentrenados para hacer un análisis de sentimiento e ironía con el objetivo de detectar las variaciones en el tono emocional asociado a diferentes términos. Por último, se emplean algoritmos clasificadores de vanguardia para extraer conclusiones acerca de las disparidades entre categorías.

En este estudio se espera descubrir patrones subyacentes, temas pevalentes y estrategias persuasivas que contribuyen a la difusión del movimiento anti-vacunas. Se espera que los hallazgos encontrados puedan servir para futuras investigaciones, discusiones políticas e intervenciones específicas destinadas a abordar y combatir este problema.

Palabras clave: vacuna, inmunización, red social, twitter, reticencia a la vacunación, COVID-19, análisis de datos, NLP, análisis del sentimiento, anti vacunas, pro vacunas.

Abstract

This project presents an analysis of anti-vaccine discourse on the social media platform Twitter using Natural Language Processing (NLP) techniques. Vaccines are biomedical technologies that stimulate the immune system to generate a protective response against specific diseases. However, despite their well-established efficacy throughout history, they generate distrust and fear in a portion of the populations.

To investigate this movement, we collect anti-vaccine messages from Twitter. Social media platforms provide a valuable source for conducting Natural Language Processing (NLP) analysis due to the abundance of data and diverse perspectives shared in real time. An exploratory investigation is carried out to identify and understand how these communities communicate on Twitter. Subsequently, a comprehensive dataset comprising tweets from recent years is collected and carefully curated for further analysis.

Firstly, text preprocessing is performed to ensure the quality and integrity of the dataset. Next, advanced NLP techniques are employed to uncover linguistic patterns and structures within the collected tweets. These techniques include n-gram extraction and temporal vocabulary analysis to identify changes in language use over time. Furthermore, a comparative analysis is conducted to obtain differences in perspectives and arguments between anti-vaccine and pro-vaccine individuals. In order to delve deeper, a pre-trained language model is utilized to perform sentiment and irony analysis, aiming to detect variations in the emotional tone associated with different terms. Lastly, state-of-the-art classifier algorithms are employed to draw conclusions regarding the disparities between categories.

This research seeks to uncover underlying patterns, prevalent themes, and persuasive strategies that contribute to the dissemination of the anti-vaccine movement. The findings of this study are expected to serve as a foundation for further research, policy discussions, and targeted interventions aimed at addressing and combating this issue.

Keywords: vaccine, immunization, social media, twitter, vaccine hesitancy, COVID-19, data analysis, NLP, sentiment analysis, anti-vaccine, pro-vaccine.

Agradecimientos

Me gustaría aprovechar este espacio para expresar mi más profundo agradecimiento a todas las personas que han estado presentes, dándome la mano en este camino que he recorrido estos últimos cinco años de mi vida.

A mi madre, por ser incondicional para mí, por estar ahí para animarme, comprenderme, apoyarme y auparme en cada momento de mi vida. Sin duda has sido mi roca y mi guía.

A mi padre, por todos los esfuerzos y sacrificios que has hecho para asegurarte de que nunca me faltara nada.

A mis amigas de la carrera: María, Bea, Anap, Barato, Ana, Aroa, Quintana, Lucía y Paloma, porque hemos estado juntas en los momentos felices y también en los difíciles.

A mis amigos Pablo y Víctor, por ser las dos personas que más me aguantan y por tener paciencia conmigo. Vuestra amistad me ha aportado fortaleza y alegría.

A mis amigos del pueblo: Miguel, Osaki, Gema, Victoria y Elena. Aunque nuestros encuentros no sean tan frecuentes como desearía, cada uno de esos momentos compartidos se ha convertido en un tesoro invaluable en mi corazón.

Por último, pero no menos importante, a mi tutor Óscar Araque, por brindarme siempre su ayuda y orientación durante la realización de este trabajo.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

Social media has become a prominent platform for the rapid dissemination of information, opinions, and sentiments. Among the wide range of topics discussed, vaccines have garnered considerable attention and sparked debates. While vaccines have been widely recognized for their effectiveness in preventing infectious diseases, certain perspectives on social platforms express skepticism and resistance towards them.

Vaccination is a simple, safe and effective method of protecting individuals against harmful diseases. The fundamental principle behind vaccines is to trigger the immune system's ability to identify and remember specific pathogens like viruses or bacteria. Therefore, when a person is later exposed to the actual disease-causing organism agent, their immune system can mount a quick and robust response, effectively preventing or mitigating the severity of the illness. Currently, vaccines exist for over 20 life-threatening diseases, such as diphtheria, tetanus, pertussis, influenza, and measles, significantly contributing to longer and healthier lives. Immunization prevents 3.5-5 million deaths annually [3].

Vaccination serves two important purposes: to protect individuals and to safeguard the broader community. While some individuals may not be able to receive vaccinations due

to factors such as age, illness, or allergies, they rely on others being vaccinated to ensure their safety from vaccine-preventable diseases. That is why achieving herd immunity is important, vaccination not only benefits individuals directly but also has a profound impact on the overall health and well-being of the community.

Herd immunity occurs when a significant portion of the population becomes immune to a disease, making it difficult for the illness to spread easily. Achieving herd immunity requires a collective effort, with individuals getting vaccinated to not only shield themselves but also to create a protective barrier for those who cannot be vaccinated.

1.1.1 Vaccine Classification

There are several types of vaccines, including inactivated vaccines, live-attenuated vaccines, toxoid vaccines, mRNA vaccines, subunit, recombinant, polysaccharide, and conjugate vaccines, and viral vector vaccines [4].

Inactivated vaccines use killed germs and require multiple doses, protecting against diseases like hepatitis A, flu, polio, and rabies [4]. Currently, there are inactivated vaccines available for COVID-19, such as Covaxin developed by Bharat Biotech and Sinopharm developed by Sinopharm Group [5, 6].

Live-attenuated vaccines utilize weakened germs to elicit a strong and long-lasting immune response, covering measles, mumps, rubella, rotavirus, smallpox, chickenpox, and yellow fever. Toxoid vaccines use toxins from disease-causing germs to create immunity against specific toxins, commonly used against diphtheria and tetanus, and may require booster shots for ongoing protection [4].

In the fight against COVID-19, mRNA vaccines have emerged as a crucial player, exemplified by the Pfizer-BioNTech and Moderna vaccines. These vaccines utilize a revolutionary approach that leverages messenger RNA (mRNA) to provide instructions to cells, enabling them to produce a harmless piece of the virus known as the spike protein. This spike protein then triggers an immune response, training the body's immune system to recognize and defend against the actual virus [7].

Unlike traditional vaccines that contain weakened or inactivated viruses or proteins, mRNA vaccines operate by delivering genetic instructions rather than viral components. The mRNA travels within a protective bubble called a Lipid Nanoparticle, facilitating its smooth entry into cells. Once inside, the cells interpret the mRNA instructions and begin building proteins that match specific parts of the pathogen, known as antigens. The immune system perceives these foreign antigens as threats, activating the production of antibodies

and T-cells to combat them. This immune response prepares the body to defend against future attacks by the real virus, potentially preventing infection and illness [8].

It is noteworthy that mRNA vaccines are not entirely new, as they have been studied extensively for other diseases like influenza, Zika, rabies, and cytomegalovirus (CMV) in the past. Furthermore, mRNA technology has also shown promise in cancer research, where it has been investigated for its ability to stimulate the immune system to target cancer cells [7].

Importantly, it should be clarified that mRNA vaccines do not enter the nucleus of the cell where our DNA is located, and therefore, they cannot alter or influence our genes [9]. This aspect ensures that the genetic material delivered by mRNA vaccines remains confined to the protein production process, contributing to their safety and effectiveness [7].

Protein subunit vaccines, like Novavax COVID-19 vaccine, contain pieces (proteins) of the virus, such as the spike protein, along with an adjuvant to enhance the immune system's response. This approach has been employed for many years and has seen success. Protein subunit vaccines are currently used to prevent other diseases like whooping cough or hepatitis B [7].

Viral vector vaccines, which is the case of Sputnik v, Johnson & Johnson's Janssen and AstraZeneca COVID-19 vaccine, use modified versions of different viruses as delivery systems to instruct cells. While they have gained attention due to their role in COVID-19 vaccination, viral vector technology has been extensively studied for cancer treatments and molecular biology research [7].

1.1.2 Vaccine hesitancy historically

Resistance to vaccination is not a recent occurrence. Since the first smallpox vaccine was developed by Edward Jenner in 1796, skepticism and suspicions about vaccines and the motivations behind their use have existed [1]. Despite the undeniable success of vaccines in reducing childhood mortality rates and eradicating diseases like polio, doubts regarding their safety and effectiveness continue to persist in the modern era.

Furthermore, incidents like the Cutter Incident in 1955, where contaminated batches of the polio vaccine caused harm, have fueled distrust in the pharmaceutical industry and led to improvements in vaccine manufacturing and regulation. Similarly, the MMR (Measles, Mumps, Rubella) vaccine faced scrutiny when a study suggested a link between the vaccine and autism. Despite subsequent studies refuting this claim, the controversy had a lasting impact on vaccine uptake. Additionally, thimerosal, a preservative used in some vaccines,

became the center of controversy due to concerns about its mercury content. Although scientific consensus deemed it safe in low concentrations, its removal from vaccines and the ensuing mixed messages caused confusion and eroded trust in vaccine regulation [1].

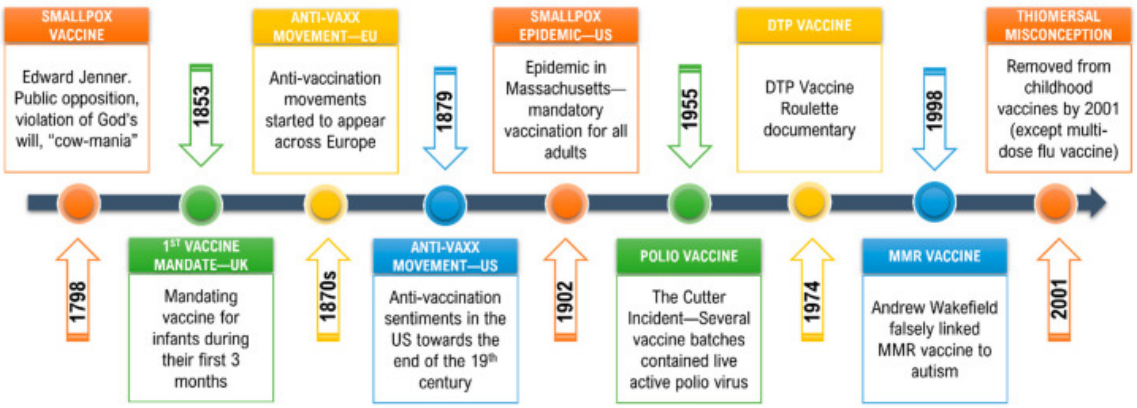

Figure 1.1: Vaccine hesitancy timeline [1].

1.1.3 Vaccine hesitancy during COVID-19 era

As of the present, in May 2023, the COVID-19 pandemic has had a profound impact worldwide. The number of confirmed cases has exceeded 700 million, with nearly 7 million reported deaths. In response to the pandemic, over 13 billion vaccine doses have been administered globally, reflecting the urgent need for effective immunization [2].

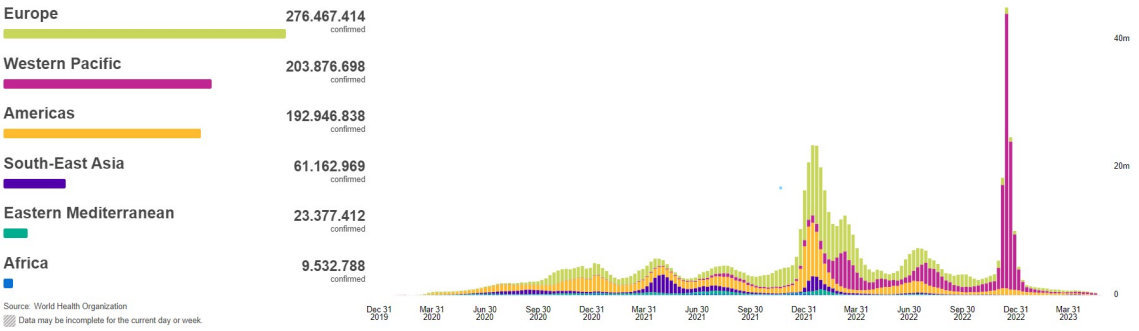

Figure 1.2: Weekly cases per region recorded by WHO [2].

The prominence of vaccines has risen significantly as a result of the COVID-19 pandemic, generating widespread interest and awareness. The pressing demand for successful vaccines to combat the virus has expedited efforts in vaccine development and distribution, capturing

unprecedented public attention towards immunization.

While COVID-19 vaccines were developed swiftly, extensive measures have been taken to ensure their safety and effectiveness. The process involves rigorous steps such as vaccine development, clinical trials, authorization or approval from regulatory bodies like the U.S. Food and Drug Administration (FDA), and the development and approval of vaccine recommendations through organizations like the Advisory Committee on Immunization Practices (ACIP) and Centers for Disease Control and Prevention (CDC). Monitoring systems are in place to ensure the ongoing safety of COVID-19 vaccines as they are distributed beyond clinical trials [7].

In spite of the robust scientific processes and the urgent need for vaccination, a complex mix of social, political, and psychological influences contributes to vaccine hesitancy [10]:

1. Individuals who perceive a lower risk of infection or exhibit apathy towards COVID-19 due to a lack of symptoms or minimal fear are more prone to vaccine hesitancy.

2. Apprehensions regarding vaccine safety, effectiveness, and potential adverse effects, as well as concerns about the potential interference with existing health conditions, contribute to vaccine hesitancy.

3. A lack of trust in governmental information and medical establishments also fuels vaccine reluctance.

4. Skepticism surrounding the accelerated development process of COVID-19 vaccines, encompassing concerns over insufficient trial durations, safety issues, and mistrust in vaccine manufacturers and policymakers, may lead individuals to decline or postpone vaccination.

5. A prevailing lack of faith in science and scientists can significantly influence individuals' perceptions of vaccines.

6. The presence of conspiracy beliefs, such as those pertaining to the origins of COVID-19 or pre-planned pandemics, further exacerbates hesitancy.

1.1.4 Vaccine hesitancy in social media

Information and knowledge about vaccines play a crucial role in vaccine hesitancy. Social media platforms like Twitter are a major source of information, and individuals resistant to vaccination rely less on authoritative sources. The lack of communication barriers on

social media allows fringe groups to spread their views easily. Misinformation gains traction not because it is considered credible, but because the potential consequences, if true, are perceived as horrifying. Consequently, there is an incentive for the dissemination of extreme propaganda, leading to an escalation of perceived threats and public fear [11].

1.1.5 Proposed solution

This project delves into the realm of Natural Language Processing (NLP) to explore anti-vaccine discourse on Twitter. By analyzing the content, language, and sentiment of these tweets, we aim to shed light on the underlying narratives, themes, and persuasive techniques employed by anti-vaccine proponents.

Throughout this project, we will navigate the landscape of NLP methodologies, including data preprocessing, feature extraction, and model training. We will employ state-of-the-art machine learning algorithms to classify tweets, detect sentiment, and uncover prevalent themes within the anti-vaccine discourse.

This project holds significant importance as it has the potential to inform public health initiatives, policy-making, and communication strategies by providing evidence-based insights into the underlying factors that contribute to vaccine hesitancy. Understanding the language and rhetoric of anti-vaccine tweets can help identify misinformation, debunk myths, and design targeted interventions to address concerns and misinformation surrounding vaccines more effectively.

1.2 Project goals

The objectives of the project are the following:

G1 Gather tweets related to vaccines from both pro-vaccine and anti-vaccine individuals and generate a comprehensive dataset.

G2 Perform a natural language processing (NLP) analysis on the collected tweets to identify common patterns, themes, and trends.

G3 Utilize a classification model, to automatically distinguish between the categories of the database, aiming to gain a better understanding of the language and distinctive features of each stance.

1.3 Structure of this document

In this section we provide a brief overview of the chapters included in this document. The structure is as follows:

Chapter 1 The introduction provides an overview of the vaccine context, vaccine hesitancy, its origins and motives, and its connection to social media. It also introduces NLP as a proposed solution.

Chapter 2 In this chapter, we introduce the enabling technologies that have been employed to facilitate the realization of this project.

Chapter 3 This chapter outlines the methodology used to capture the data use to collect anti-vaccine and pro-vaccine tweets.

Chapter 4 In this part, the collected tweets are processed and analyzed. The analysis includes temporal analysis of n-grams, frequency differences between categories, sentiment analysis, and hashtag analysis.

Chapter 5 This chapter explores the use of two classifiers to evaluate the difficulty of classifying tweets into categories. The goal is to assess the vocabulary differences between the categories.

Chapter 6 The final chapter summarizes the achieved goals, discusses encountered challenges, and provides conclusions drawn from the analysis. It also suggests potential areas for future research.

CHAPTER 2

Enabling Technologies

In this section, the enabling technologies utilized in the development of this project will be introduced. These encompass a variety of powerful programming tools and libraries that formed the foundation for conducting NLP analysis and extracting valuable information about the anti-vaccine movement on Twitter.

2.1 Programming and development environment Technologies

2.1.1 Python

Python is a popular high-level programming language that is widely used in a variety of applications, including web development, scientific computing, data analysis, artificial intelligence, and automation.

Python is known for its simplicity, readability, and flexibility, which make it easy to learn and use. It features a vast standard library and a large number of packages that can be easily installed using package managers like pip. Python supports various programming paradigms, including procedural, object-oriented, and functional programming, and it has a large and active community that contributes to its development and maintenance.

Python is a popular choice for data science due to its rich ecosystem of data manipulation and analysis libraries, such as NumPy, Pandas, and Scikit-learn. These libraries enable users to easily process, visualize, and model data.

2.1.2 Jupyter Notebooks

Jupyter Notebook is an interactive computational environment that allows users to create and share documents that combine live code, equations, visualizations, and explanatory text. It is widely used in data science, scientific computing, and machine learning because it enables users to easily explore and manipulate data, experiment with different algorithms, and document their findings. Jupyter Notebook supports many programming languages, including Python, R, Julia, and MATLAB, and it can be run on a local machine or in the cloud. Its interface is web-based and provides a flexible and intuitive way to work with data and code.

For this project, we utilized Jupyter Notebook as our coding environment to write and execute Python code. Jupyter Notebook provided us with an interactive and flexible platform to explore and analyze our data, and to document our analysis through a combination of code, visualizations, and text.

2.2 Data acquisition technologies

2.2.1 Snscrape

Snscrape [12] is a scraper for social networking services (SNS). It facilitates the extraction of data from Twitter, specifically tweets, in a simple and efficient manner. snscrape emulates the behavior of a real user accessing Twitter through a web browser. By doing so, it gains access to public Twitter data without additional authentication, access restrictions or request limits.

2.3 Natural language processing technologies

2.3.1 NLTK

NLTK (Natural Language Toolkit) is a popular Python library for natural language processing. It provides a wide range of tools for processing and analyzing textual data, such

as tokenization, stemming, lemmatization, and part-of-speech tagging. By using NLTK's tools, we were able to effectively preprocess the text and extract meaningful insights from it.

2.3.2 Gensim

Gensim [13] is a Python library for topic modeling, document similarity, and text processing. It provides a set of tools for analyzing large collections of textual data, including methods for preprocessing, vectorization, and modeling.

In this project, we employed Gensim's Phrases tool to automatically detect and create bigrams and trigrams from the input text. Phrases apply statistical methods to identify frequently co-occurring terms and merge them into single phrases, which can help improve the accuracy and relevance of downstream analyses.

2.4 Sentiment analysis technologies

2.4.1 Transformers

Transformers is a popular open-source library for natural language processing (NLP) that provides a wide range of pre-trained models and tools for various NLP tasks. It is based on the state-of-the-art transformer architecture and has been used to achieve state-of-the-art performance in many NLP benchmarks.

In this project, we used Transformers for Python to conduct sentiment and irony analysis on textual data. Specifically, we utilized models based on the Twitter-RoBERTa-base, which is a pre-trained language model that is based on the RoBERTa architecture, which itself is an extension of the BERT (Bidirectional Encoder Representations from Transformers). It is trained on over 58M tweets and finetuned for sentiment analysis with the TweetEval benchmark. The RoBERTa models generally outperform the other models across various tasks, including classifying tweets for Emoji, Emotion, Hate, Irony, Offensive language, Sentiment, and Stance [14].

By using Transformers for Python, we were able to easily load and utilize a high-quality pre-trained model for sentiment analysis, without the need for extensive training or fine-tuning. This allowed us to perform accurate and efficient sentiment analysis on the input text, which in turn helped us gain insights into the attitudes and opinions expressed in the data.

2.5 Data modelling technologies

2.5.1 Numpy

NumPy is a Python library for numerical computing. It provides a powerful array object, called ndarray, which efficiently stores and handles large arrays of homogeneous data. NumPy's efficient handling of large arrays and matrices makes it an invaluable tool for data processing tasks.

One of NumPy's key strengths is its extensive collection of mathematical functions designed for performing mathematical operations on arrays. These functions enable you to carry out statistical calculations, such as means, variances, and correlations. In addition, NumPy also integrates well with other libraries commonly used in data processing workflows, such as Pandas for data manipulation and Matplotlib for visualization.

2.5.2 Pandas

Pandas is a popular open-source data analysis and manipulation library for Python. It offers an extensive range of tools and functions specifically designed to handle structured data, including tables and time series.

One of the key data structures in Pandas is the DataFrame, which is a two-dimensional table-like data structure that is similar to a spreadsheet or a SQL table. DataFrames in Pandas provide a robust and flexible way to manipulate and analyze data. They can be created from a variety of data sources, including CSV and Excel files, SQL databases, and Python lists and dictionaries. Once created, DataFrames can be manipulated and transformed in a variety of ways, including selecting, filtering, and aggregating data.

DataFrames also support a variety of operations for handling missing data, such as filling in missing values or dropping rows with missing data. In addition, DataFrames can be merged and joined with other DataFrames to combine data from different sources.

2.5.3 GSITK

GSITK [15, 16] is a library that is built on top of scikit-learn, and it serves as a valuable tool for NLP machine learning projects. It provides a variety of tools for processing and analyzing social media data.

2.6 Data visualization technologies

2.6.1 Matplotlib

Matplotlib is a popular open-source data visualization library for Python. It provides a wide range of functions for creating various types of plots and charts, such as line plots, scatter plots, bar charts, and histograms. Matplotlib is widely used in data science, engineering, finance, and many other fields where data visualization is important.

2.6.2 Scattertext

Scattertext [17] is a Python library designed to visualize and analyze the differences in language between multiple categories of texts. It serves as a tool for identifying distinctive terms within corpora and presenting them in an interactive HTML scatter plot.

2.6.3 Shifterator

Shifterator [18] is Python library that allows comparing the relative frequencies of certain types between two systems. Specifically, it calculates the proportion of each type in the two systems and compares the difference in proportions to assess whether there is a shift in the usage of that type from one system to the other. It can be used to detect language variation and change over time, among other things. In this project, we employed the ProportionShift class to conduct a comparative analysis of the relative frequencies of diverse types of vaccine-related language.

2.6.4 Plotly

Plotly [19] is a Python library that allows you to create interactive and dynamic graphs. It provides a wide range of visualizations, including line charts, scatter plots, bar charts, and more. The library can be used for both exploratory data analysis and data visualization in reports or presentations.

Plotly was a valuable tool in this project because we needed to display a large amount of information in a single graph, which would have been confusing without any interactivity. Its interactive features allowed users to zoom in and out, hover over data points for more information, and explore the data more deeply.

2.7 Machine Learning technologies

2.7.1 Scikit-learn

Scikit-learn is an open-source Python library for machine learning and statistical modeling that offers efficient tools for various tasks like classification, regression and clustering. It provides a consistent API for different machine learning algorithms, allowing easy switching between models and evaluating their performance. In this project, we used Scikit-learn to extract features from text data, split the data for training and testing, and train two classifiers, Linear SVC (Linear Support Vector Classification) and Random Forest, for the purpose of text classification.

CHAPTER 3

Dataset capture

3.1 Introduction

This section covers the collection process of the tweets for the dataset generation. The objective was to gather English tweets from December 2020 to the present, from individuals expressing both anti-vaccine and pro-vaccine sentiments. The decision to start from December 2020 is because it marks the beginning of COVID-19 vaccination campaigns [20].

3.2 Data acquisition

Snscrape was used as the tool of choice to extract tweets for the study. By leveraging the capabilities of snscrape, relevant tweets were collected based on specific search criteria and parameters.

To ensure accuracy and minimize errors, we conducted a thorough and detailed review of individual tweets, instead of extracting a large volume of tweets massively. When encountering posts associated with the anti-vaccine stance, we did not save the tweet itself, but rather recorded the corresponding username. Subsequently, we extracted multiple tweets

from each user categorized as "anti_vaccines", which significantly expanded our dataset. This approach enabled us to obtain a substantial volume of messages while only manually reviewing a small percentage of them.

To identify relevant terms for collecting vaccine-related tweets and categorizing them as anti-vaccine or pro-vaccine, a study was conducted on the communication patterns of both anti-vaccine and pro-vaccine groups on Twitter. Obvious terms like "vaccine" were not very useful as they could belong to either category. Another term, "anti-vaxxer", an informal synonym for "vaccine skeptic", resulted in many pro-vaccine tweets criticizing anti-vaccine individuals, as well as tweets from vaccine-hesitant users, which made it less useful. However, a term frequently used exclusively by anti-vaccine individuals was "jab", used as a synonym for vaccines.

By searching for phrases like "kill jab", "covid jab", and "death jab" between December 2020 and the present, a majority of the tweets were found to be against COVID-19 vaccines. It's worth noting that beyond these dates, "jab" is not used as frequently for vaccines but more commonly in the context of a sting or video games.

A manual search was performed for users expressing anti-vaccine sentiments using the terms "death jab" and "kill jab" month by month until a consistent number of users were obtained. However, after collecting around 350 users and categorizing them as "anti_vaccines", it became apparent that there might be differences in language usage between individuals who have always been vaccine skeptics and those who have received vaccines in the past but are now against the newly emerged COVID-19 vaccines. To address this, a distinction was made in the search for anti-vaccine tweets, differentiating between two categories:

1. "Anti_all_vaccines": This category includes tweets from individuals who have seemingly been against vaccines other than COVID-19 vaccines.

2. "Anti_covid_vaccines": This category comprises tweets from individuals specifically against COVID-19 vaccines.

In addition, a third category was created:

"Pro_vaccines": This category consists of tweets from individuals expressing support for vaccines.

3.2.1 Anti_all_vaccines

For the "anti_all_vaccines" category, the search focused on finding anti-vaccine users before the existence of COVID-19 in late October 2019 [21]. Several terms were searched and the following ones yielded a significant number of tweets:

- #vaccineskill
- vaccines poison
- vaccine injured
- #VaccinesAreNotTheAnswer
- #avoidAllVaccines
- #vaccinesCauseAutism

The searches were conducted from October 30, 2019, backward, month by month, until approximately 300 users were obtained.

3.2.2 Anti_covid_vaccines

The search for this category focused on identifying individuals who specifically opposed COVID-19 vaccines and did not express opposition towards other vaccines. To achieve this, tweets were searched using the keyword "real vaccine" with the intention of finding messages that explicitly stated that COVID-19 vaccines were not considered comparable to previous vaccines like, for example, the polio vaccine. Figure 3.1 provides an example.

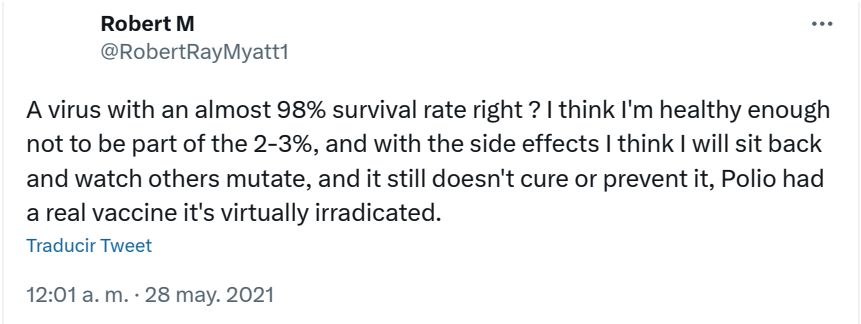

Figure 3.1: Example of searched tweets.

With this search criteria, it cannot be guaranteed that the included users are in favor of all vaccines except for COVID vaccines. It simply indicates that they express a greater

willingness to receive traditional vaccines compared to the ones developed specifically for COVID-19. A total of 150 users were gathered between 2023 and January 1, 2021, and an additional 150 users were collected between January 1, 2021, and December 31, 2021, for the "anti_covid_vaccines" category.

3.2.3 Pro_vaccines

For the "pro_vaccines" category, it was relatively easier to search for tweets with pro-vaccine hashtags, such as #vaccinessavelives and #vaccineswork. Some tweets were also found using the terms "fully vaccinated" and "vaccines save lives". Nevertheless, for other vocabulary related to supporting vaccines, it was difficult to determine if they were used sarcastically. For example, phrases like "glad to be vaccinated" or "thankful to be vaccinated" were commonly used, but they were also employed sarcastically by some anti-vaccine individuals. These individuals mock the fact that these vaccines do not guarantee 100 % protection against COVID-19 but rather provide milder symptoms in case of infection.

Similar to the previous category, we collected 230 users between December 2020 and January 1, 2021, and an additional 250 users between January 1, 2021, and the present (February 2023). Since this category takes a pro-vaccine stance, in contrast to the other three categories which are anti-vaccine, it seems reasonable to search for more users within this particular category.

3.2.4 Tweet collection metodology

The procedure employed to gather tweets from users in each category followed these steps:

1. Identification of relevant keywords: After reviewing a significant number of tweets, the most frequent terms were identified. These terms included: **#vaccineskill, #vaccinessavelives, #vaccineswork, anti vaxxer, AstraZeneca, Covaxin, flu shot, gene therapy, herd immunity, jab, Moderna, mRNA, Novavax, Pfizer, Sputnik, unvaccinated, vaccine, vax.**

2. Searching for tweets per user: For each user in every category, a search was conducted using the specified keywords. The search was conducted within two date ranges: from December 31, 2020, to January 1, 2021, and from January 1, 2021, to February 2023. In each time period, a maximum of 10 tweets per keyword were sought from each user.

3. Removal of duplicate tweets: Any duplicate tweets within the collected dataset were eliminated to ensure data integrity.

4. Manual review and removal of spam users: A manual review of the users who had the highest number of tweets was conducted to identify and eliminate any accounts engaged in spamming activities.

We collected and stored the tweets in a Pandas DataFrame, it was structured as shown in Figure 3.2 and the resulting distribution of users and tweets in each category is summarized in Table 3.1.

	Datetime	Tweet Id	Content	Username	Type
0	2020-12-01 00:11:51+00:00	1333564431679229952	2/2\nIf I was a doctor, I'd probably read @The...	AnGobanSaor	pro_vaccines
1	2020-12-01 02:52:37+00:00	1333604887540637698	Abolish National Vaccine Injury Act &a...	Poet_Carl_Watts	anti_all_vaccines
2	2020-12-01 06:34:34+00:00	1333660743531593728	@BusyDrT Covid vaccines are the mark, they wil...	whizzle_phizzle	anti_all_vaccines
3	2020-12-01 08:56:37+00:00	1333696492687687681	@LLinWood @suzydymna @BrianKempGA Am I the onl...	realArizonaDan	anti_covid_vaccines
4	2020-12-01 11:40:41+00:00	1333737783123349506	@leanansidh3 @hotboxwitgod @GovofCO Actually v	jabbingmesoftly	anti_all_vaccines

Figure 3.2: DataFrame Structure: Column Distribution Overview.

Category	**Tweet Count**	**User Count**
anti_all_vaccines	11,794	224
anti_covid_vaccines	13,560	297
anti_vaccines	13,688	340
pro_vaccines	15,024	411

Table 3.1: Number of users and tweets per category.

Snscrape follows a reverse search approach, starting from the specified end date and moving backwards in time. In many cases, it was able to find the 10 tweets very close to the end date. Consequently, as the search progresses backward in time, the number of collected tweets decreases. For that reason, the search was conducted in two temporal periods. This was necessary because without this approach, there would have been very few tweets available for dates close to December 2020 [20]. This particular date is significant as it marks the beginning of COVID-19 vaccination, and it could provide valuable insights and information. A graphical representation illustrating the monthly distribution of tweets across different categories is presented in Figure 3.3.

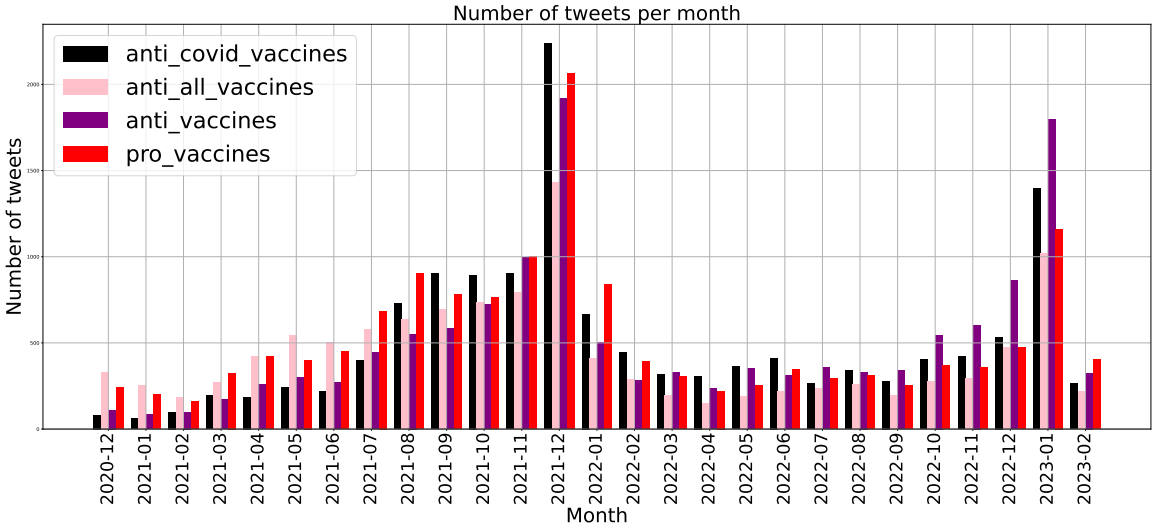

Figure 3.3: Monthly tweet count per category.

CHAPTER 4

Data analysis

In this chapter, we utilize Natural Language Processing (NLP) techniques to perform a thorough analysis of language patterns in the anti-vaccine and pro-vaccine messages on Twitter. Our analysis encompasses multiple stages, including data preprocessing, feature extraction, exploration of temporal patterns, comparison of frequencies across categories, sentiment assessment, and analysis of hashtags. Additionally, we provide valuable contextual information that enhances the understanding and interpretation of our data analysis results, offering a comprehensive perspective on the vaccine discussion within the Twitter community.

4.1 Preprocess and n-gram detection

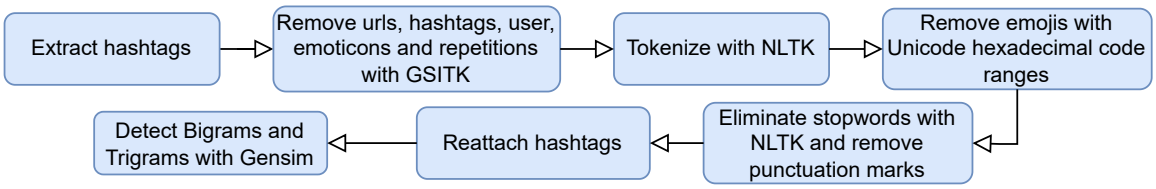

Figure 4.1: Preprocess pipeline.

The GSITK library was utilized to perform various transformations on the text. This module offered various funcionalities:

- Replaces URLs starting with "http://", "https://", or "www." with the tag "<url>".
- Inserts spaces around forward slashes ("/") to separate them from other characters.
- Replaces Twitter usernames with the tag "<user>".
- Converts text to lowercase and adds the "<allcaps>" label to indicate that the original text was in all uppercase letters.
- Replaces hashtags' pound sign with the "<hashtag>" tag.
- Substitutes numbers with the "<number>" tag.
- Removes select emoticons and labels them with "<smile>".
- Replaces sequences of repeated punctuation marks with a singles occurrence followed by the tag "<repeat>".
- Identifies words with repeated characters and appends the tag "<elong>" to the word.

Some modifications were made to fulfill the specific requirements of text processing. The main aim was to avoid retaining tags with $< >$ symbols in the processed text. The allcaps functionality was modified to prevent unnecessary word divisions.

Then, the text was splitted into tokens using the TweetTokenizer class from the NLTK library. Subsequently, emojis were removed using regular expressions that represent the Unicode hexadecimal code ranges of emojis. The emoji module was also utilized to detect and eliminate any remaining emojis.

In order to eliminate stopwords from the text, we utilized the "StopWordsRemover" module provided by GSITK, as well as the stopwords list from NLTK. Additionally, punctuation marks were also removed.

Lastly, an issue was addressed regarding hashtags that contained numbers, such as "#ivax2protect". In order to handle this, all hashtags were extracted from the original text before preprocessing, stripped of pound signs (#), and converted to lowercase. After the text was processed, the hashtags were reattached to each tweet, preserving their original form.

To detect common n-grams, sequence of n consecutive words, I used a module of the Python library Gensim: "gensim.phrases", which detect bigrams from a stream of sentences

with the function "Phrases". The bigram and trigram models are trained with a minimum count of 5 and a threshold of 10. By setting min_count=5, only bigrams that occur at least 5 times in the text will be considered. This helps filter out less frequent bigrams and reduces noise in the results. With a threshold value of 10, a relatively high association score is required to form bigrams. A threshold value of 10 means that only bigrams with a stronger association, occurring much more frequently than expected by chance, will be considered. This helps filter out bigrams with weak or spurious associations.

As a result, a more precise and meaningful selection of bigrams is obtained. The total number of generated bigrams will be lower compared to more flexible configurations, but the selected bigrams will be more reliable and representative of relevant patterns in the text. To obtain the trigrams, the Phrases function is applied once again to the previously generated list of bigrams.

4.2 Vocabulary evolution

For these graphs, the top 5 n-grams with the highest count were obtained for each month. The frequency of each term was then calculated as a percentage by dividing its count by the total number of n-grams in that month. To enhance the visualization of the extensive information, the Plotly library was utilized, which provides the capability to create interactive graphs. However, in this document, we have included static versions of these graphs, which show only a partial legend. For the complete legend, refer to Appendix C. To provide additional temporal context, we added vertical lines to indicate relevant events during each respective time period [22, 23].

Focusing on anti vaccines categories, "herd immunity" began as a very common term in December 2020 when COVID vaccines were first being administered, but as the months passed, it became less and less frequently used as shown in Figures 4.2 and 4.3.

In Figures 4.4 and 4.5, it also appears as part of the trigram "natural herd immunity" which gives a little bit of context of how it is used. This is because many people initially believed that achieving herd immunity required becoming infected and developing natural immunity. But, WHO has declared that herd immunity should be achieved through vaccination rather than exposing people to the virus [24].

Additionally, individuals may feel less fearful of the virus because they perceive a high likelihood of recovering from infection. For example, in countries like Spain or the US, COVID-19 has shown a relatively high recovery rate [25]. This notion is reflected in Figures

4.4 and 4.5, where we observe the frequent occurrence of the term "virus survival rate" during early 2021. The term "virus survival rate" refers to the chances of surviving after being infected by a specific virus. Consequently, some individuals may question the necessity of getting vaccinated, believing that natural infection alone will offer adequate protection.

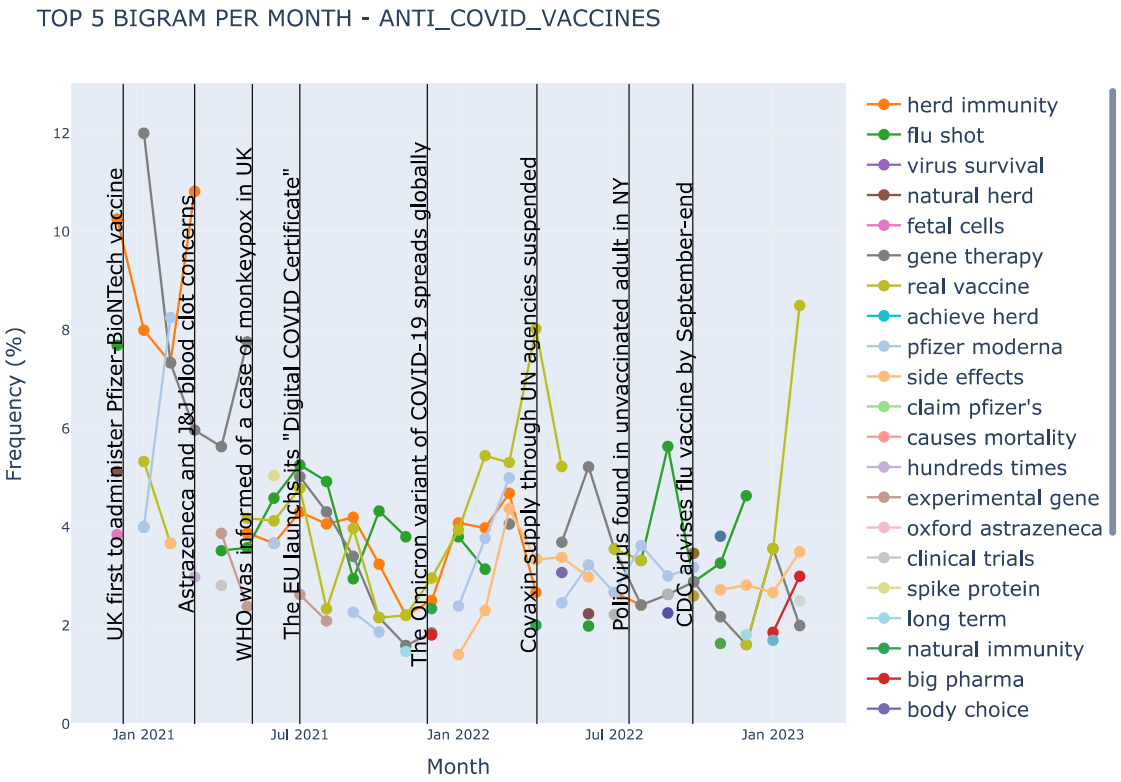

Figure 4.2: Evolution of bigrams in the "anti_covid_vaccines" category.

Another concept that starts off high in December 2020 and gradually decreases is "gene therapy" (Fig. 4.2, 4.3). Gene therapies encompass deliberate alterations made to an individual's DNA with the aim of curing or alleviating a genetic disorder [9]. The attention on gene therapy is due to the lack of knowledge about what it actually means for a vaccine to be made of mRNA, which generated collective hysteria, as gene therapy does modify DNA, while mRNA vaccines do not. As a result, conspiracy theories started to spread, suggesting that mRNA vaccines were being used by those in power to modify DNA for purposes such as controlling humanity. This technology also appears in the trigram "experimental gene therapy" and remains one of the most frequent trigrams, particularly in the "anti_all_vaccines" category (Fig. 4.5).

Related to the discussion about the experimental nature of the vaccine, anti-vaccine users often express concerns about the rapid development of COVID-19 vaccines. This

has generated fears that the vaccines may not have undergone sufficient testing. As a result, "clinical trials" appears in Figures 4.2 and 4.3 as a top bigram in April, 2021 and it also mentioned as "still clinical trials" in June 2021, for anti_all_vaccines tweets. It can be observed that this concern persists over time, as "never tested" appears as a relevant bigram in October 2021 for "anti_covid_vaccines" (Fig.C.1).

Linked to the mistrust in the testing conducted for the approval of these vaccines, there is concern about the potential side effects they may cause. "Side effects" emerges as a trend for both types of anti-vaccine groups (Fig. 4.3, 4.3) from mid-2022 onwards and in the following months. Another relevant bigram referring to the same issue is "adverse reactions", which appears in the context of anti_all_vaccines in March 2021 (Fig. 4.3). There is particular agitation regarding potential unknown long-lasting side effects, which is reflected in the bigram "long-term effects" present in discussions about "anti_covid_vaccines" during various months throughout 2021 and 2022, as shown in Figure 4.2.

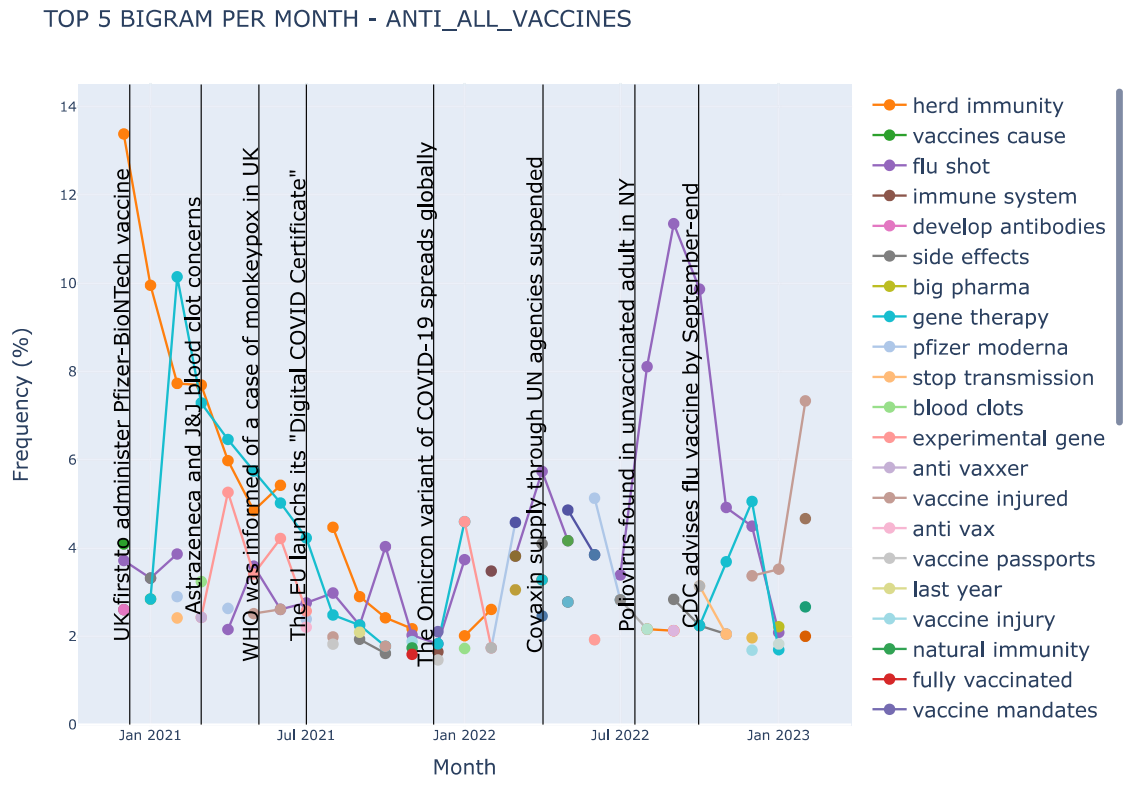

Figure 4.3: Evolution of bigrams in the "anti_all_vaccines" category.

One of the side effects that sparked significant discussion was "blood clots", which appeared for the "anti_covid_vaccine" category in March 2021 in Figure 4.2. This is due to a controversy surrounding the AstraZeneca and Johnson & Johnson COVID vaccines, as there were reports of blood clots associated with these vaccines. Although these incidents

were rare, further investigations were conducted as a precautionary measure. Consequently, there were temporary pauses in the administration of Johnson & Johnson and AstraZeneca vaccines [26, 27]. Therefore, we also observe the occurrence of the bigram "Oxford AstraZeneca" during that particular month in Figure 4.2.

On a different note, it is noteworthy that starting from the second half of 2021, there was significant attention given to COVID passports. These passports serve as regulatory measures implemented to monitor international travel and mitigate the spread of the COVID virus between countries. A significant development in this regard was the introduction of the COVID certificate by the European Commission on July 1, 2021 [28]. Thus, the term "vaccine passports" gained prominence as a trending topic in August 2021 as seen in Figure 4.3.

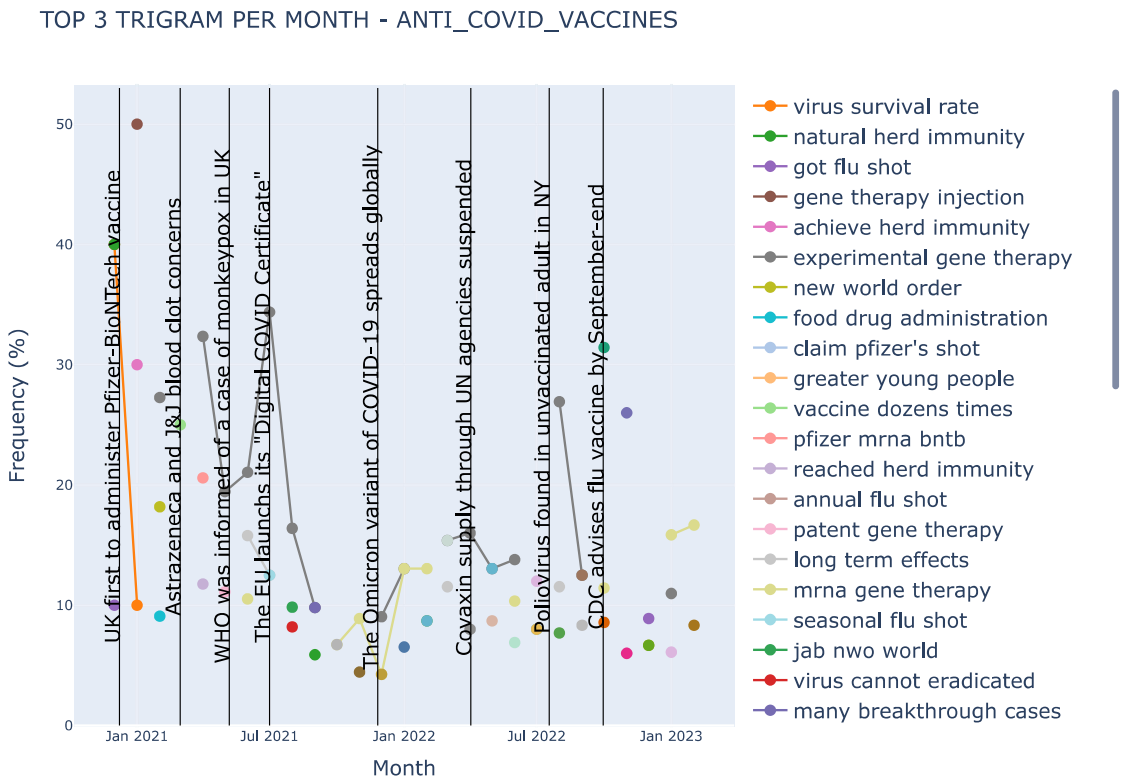

Figure 4.4: Evolution of trigrams in the "anti_covid_vaccines" category.

While side effects of the AstraZeneca and Johnson & Johnson vaccines were a cause for concern in March, there was less discussion about them later on. On the other hand, the Pfizer brand remained a subject of controversy in both anti-vaxxer categories. Pfizer and Moderna were the first vaccines to be administered and are highly mentioned, especially Pfizer. In the following months, Moderna appears less frequently, mainly as a frequent

bigram associated with Pfizer, as seen in the bigram "Pfizer Moderna" in Figures 4.2 and 4.3.

In March 2022, Pfizer was involved in false rumors [29], resulting in the term "Pfizer documents" appearing in the anti_all_vaccines category (Fig. C.3) and "read Pfizer doc" in the "anti_covid_vaccine" trigram graph (Fig. C.1). These rumors claimed that Pfizer released documents revealing an efficacy of 12% and that Pfizer refused to provide detailed reports until 2055. However, it was concluded that these rumors stemmed from a misinterpretation of data and facts. This example illustrates how misinformation and disseminating information from unreliable sources often drive anti-vaxxers.

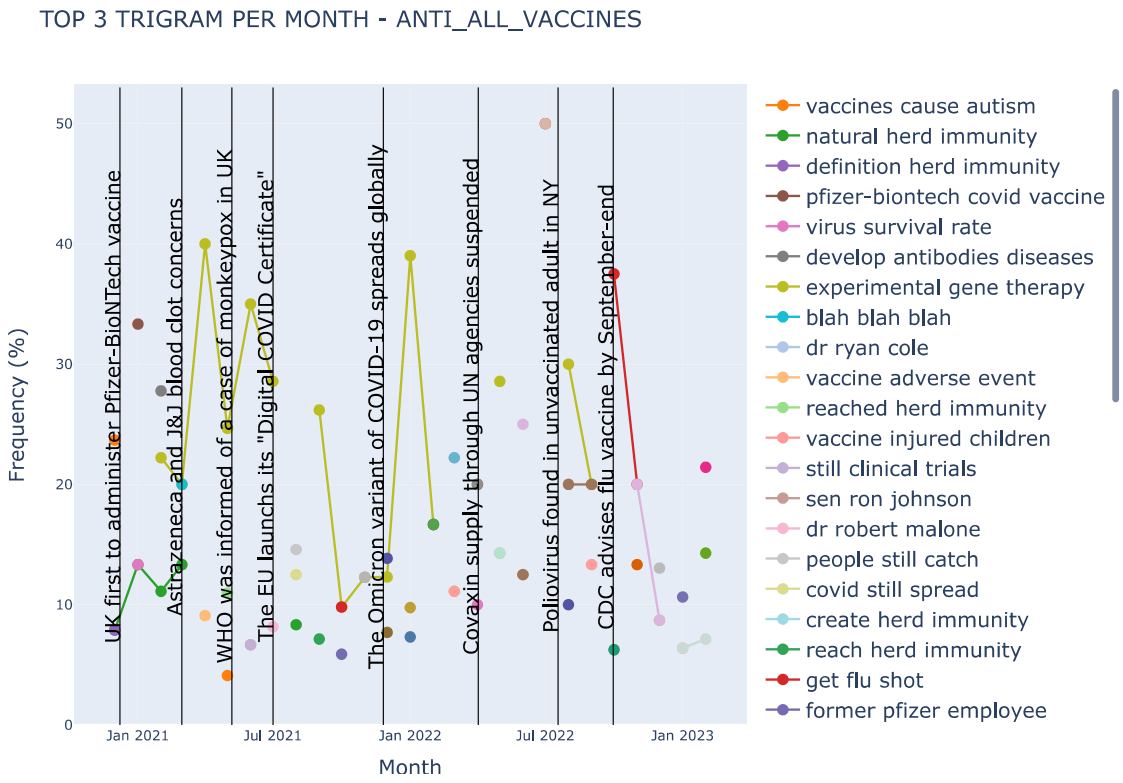

Figure 4.5: Evolution of trigrams in the "anti_all_vaccines" category.

Lastly, although "flu shot", the informal term for the influenza vaccine, appears as the most frequent bigram for many months in both categories, we can observe a significant peak in September and October 2022 for both categories, particularly in the context of "anti_all_vaccines" in Figure 4.3, where it experiences a significantly substantial increase. During these months, there is heightened discussion about the influenza vaccine as it is administered each season due to the mutating nature of the flu virus and health organizations such as the Centers for Disease Control and Prevention (CDC) recommend initiating vaccination during this time [30]. In fact, in September 2021, "flu shot" is also among the

most frequent terms, but it does not exhibit as abrupt and significant increase as observed in 2022. This could be attributed to the fact that the conversation surrounding flu shots in 2021 may have been overshadowed by discussions about COVID-19.

Moving on to "pro_vaccines" graph (Fig. 4.6), we can observe how "herd immunity", also starts off with high frequency but gradually decreases over time. Similar to the previous categories, the same pattern is seen for Pfizer and Moderna vaccines. Being the first ones administered [31], they receive significant attention initially and remain a media focus over time.

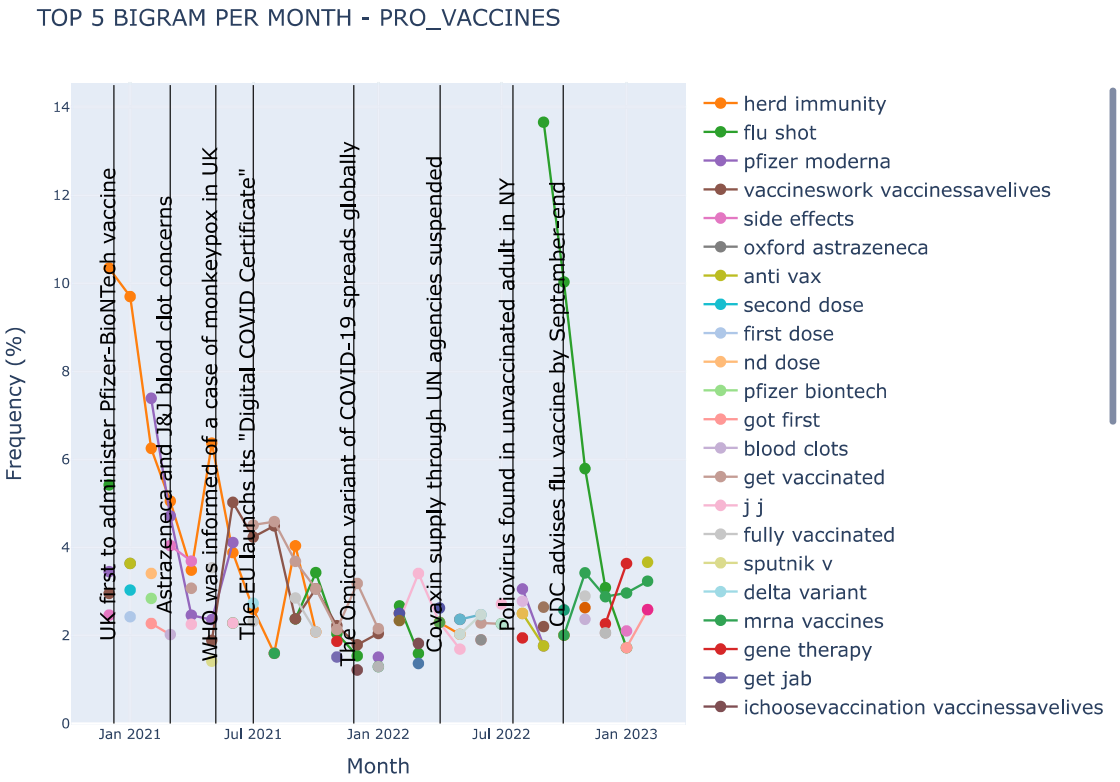

Figure 4.6: Evolution of bigrams in the "pro_vaccines" category.

However, AstraZeneca only becomes a trend in March 2021, coinciding with the aforementioned controversy. The trigram graph (Fig. 4.7) provide some insights into the context surrounding this controversy during that month, where the term "blood clots" appears as "rare blood clots", emphasizing the low likelihood of experiencing this side effect. On the other hand, Johnson & Johnson (J&J), the other vaccine involved in the controversy, is mentioned in several months throughout 2021 and 2022. Additionally, there were occurrences of the Sputnik V COVID vaccine in May 2021 as shown in Figure 4.6, which had not been previously prominent in anti-vaccine tweets.

One word that receives significant attention is “dose”, which is mentioned for several months throughout 2021 and 2022. By examining the trigram graph (Fig. 4.7), we can discern the context in which this word is used. It appears in trigrams such as “received first dose” and “second dose Pfizer”, which could indicate that the usage predominantly stems from individuals in pro-vaccine circles sharing on social media that they got vaccinated or which vaccine they have received.

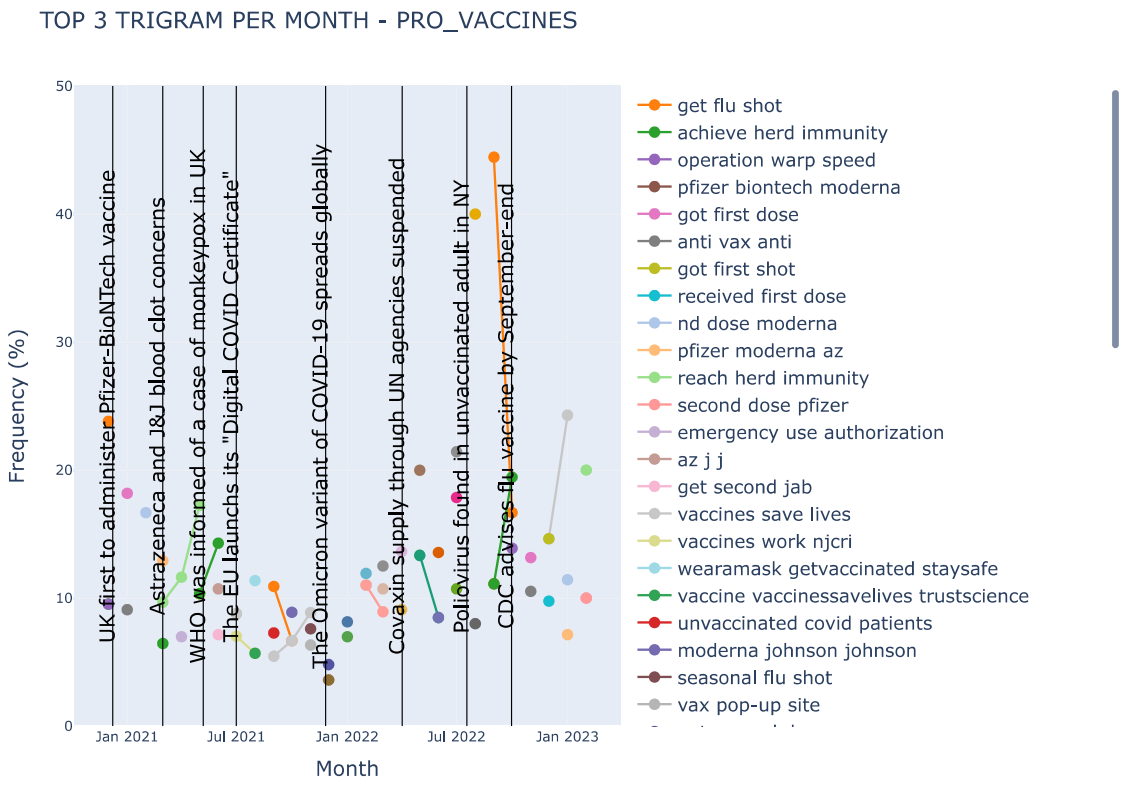

Figure 4.7: Evolution of trigrams in the “pro_vaccines” category.

In July 2021, it was declared that the Delta variant was spreading globally [32], as indicated by the appearance of the bigram “Delta variant” in Figure 4.6. This led to the introduction of boosters. These additional doses have proven effective in enhancing immunity, particularly for individuals such as the elderly and those with compromised immune systems who exhibited a weaker response to the initial doses of the vaccines. The need for these booster doses increased due to the higher transmissibility of the Delta variant [33]. Consequently, they were approved by the FDA in August 2021 [34, 35]. However, it was not until February 2022 that “booster” gained significant attention in ongoing discussions (Fig. C.5). It is interesting to note that in September 2022, “bivalent booster” emerged as a frequent bigram. This bivalent booster is an updated version of the vaccine that includes both the original strain of the virus and a strain derived from the BA.5 Omicron variant

[36].

Another difference compared to the previous graphs is the appearance of numerous hashtags encouraging the population to get vaccinated, such as #vaccinessavelives, #vaccineswork, #unite2fightcorona and #ichoosevaccination. These hashtags consistently appear throughout the dataset, almost in a timeless manner.

It is worth mentioning that the bigram "endpolio polio" appear in relation to the reemergence of poliovirus cases in the United States in July 2022 [37]. "Endpolio" is used as a hashtag in this context. Similar to the previous categories, "flu shot" appears in several months over the two-year period but experiences a significant spike in September 2022.

4.3 Vocabulary comparison

In this section, our focus is on exploring the differences among various categories within the context of vaccine discussions.

4.3.1 Shifterator

Shifterator is a tool that allows us to visualize the variations in word usage across these categories. We employ the proportion shift approach, which compares the relative frequencies of words in the first and second texts. Specifically, if $p_i^{(1)}$ represents the relative frequency of word i in the first text, and as $p_i^{(2)}$ represents its relative frequency in the second text, the proportion shift is calculated as: $\delta p_i = p_i^{(2)} - p_i^{(1)}$. If a word has a positive difference in relative frequency ($\delta p_i > 0$), it indicates that the word is relatively more common in the second text. Conversely, if the difference is negative ($\delta p_i < 0$), the word is relatively more common in the first text. By ranking the words based on these differences, we can visualize them as a word shift graph.

Let's start by analyzing the vocabulary difference between the two types of anti-vaccine groups in Figures 4.9 and 4.8. We can see that the main distinction lies in the keywords used to retrieve tweets from each category. In the "anti_all_vaccines" category, there is a high frequency of the word "injury" and its derivatives, referring to people who have suffered some adverse effects due to vaccines.

On the other hand, in the "anti_covid_vaccines" category, the term "real vaccine" appears much more frequently. As a consequence of this, for "anti_covid_vaccines" we find bigrams like "definition vaccine", "prevent transmission", "prevent infection", and "provide

immunity". These terms are used in the context of comparing the new COVID vaccines with those from the past. This is because these vaccines do not guarantee complete protection from being infected; instead, they enable individuals who are infected to experience milder symptoms. For example, the COVID vaccine can prevent the disease from progressing to a more severe respiratory condition [38]. The bigram "definition vaccine" refers to the necessity of expanding the traditional definition of a vaccine to include the newly developed ones that emerged during the COVID pandemic. This expansion has led to an increased level of resistance and rejection towards these vaccines.

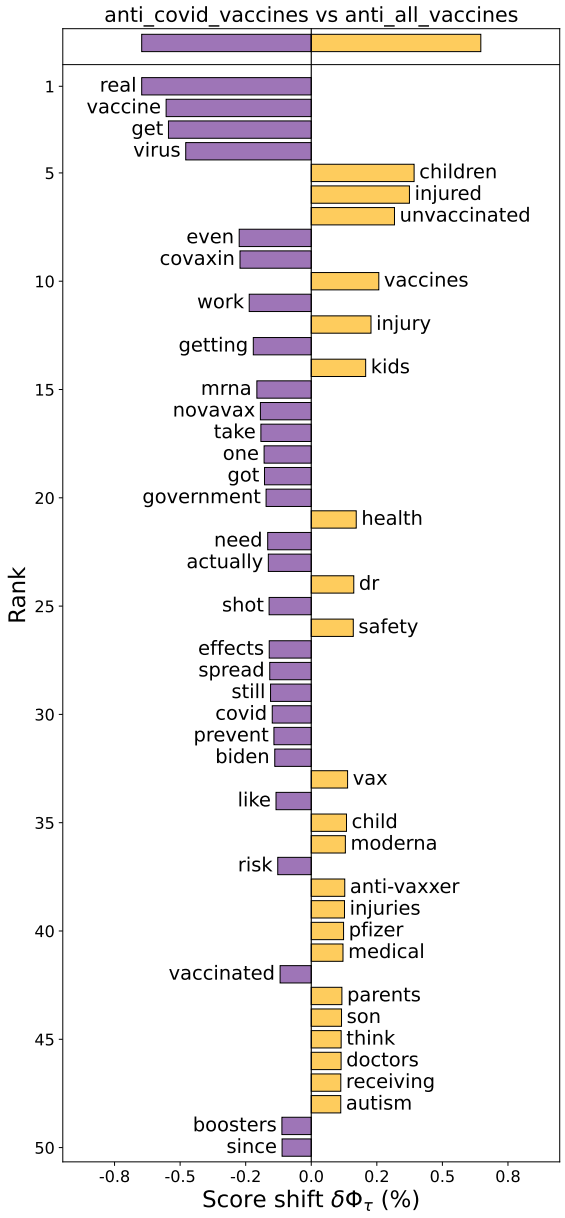

Figure 4.8: Shift: ACV vs AAV unigrams.

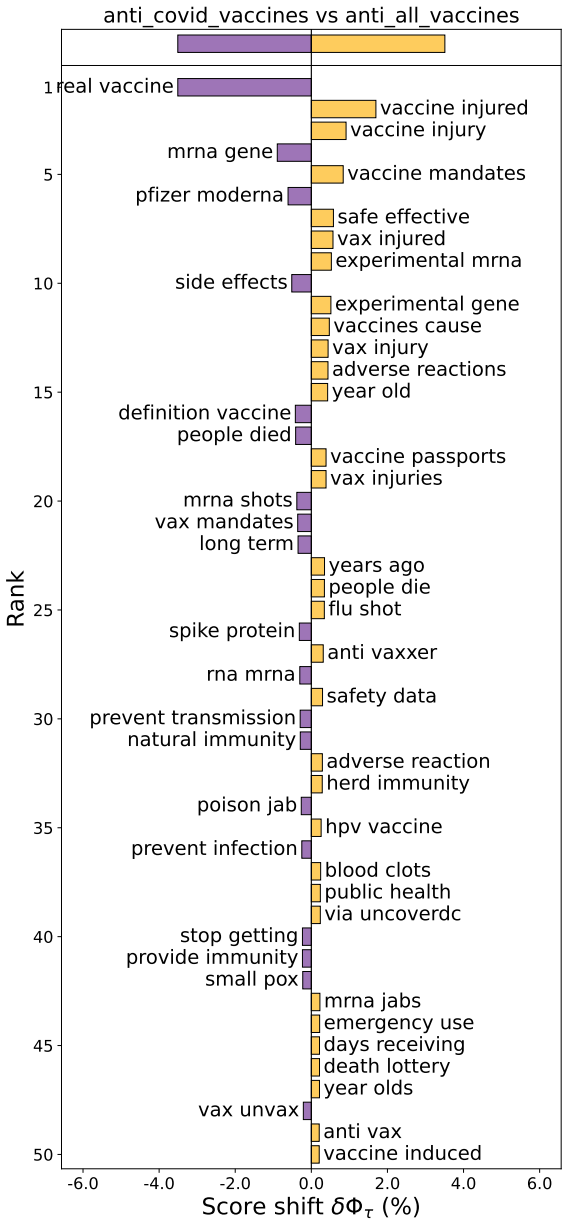

Figure 4.9: Shift: ACV vs AAV bigrams.

In contrast, in the "anti_all_vaccines" category, we observe a higher frequency of terms such as "anti-vax" and "anti-vaxxer" compared to the "anti_covid_vaccines" category. This discrepancy can be attributed to the fact that many individuals who oppose COVID vaccines do not consider them as vaccines at all, as they have previously received other vaccines without hesitation, and therefore they do not identify themselves as "anti-vaxxers" according to the conventional definition.

Within the "anti_all_vaccines" category, there is a greater emphasis on terms related to kids, such as "children" and its synonyms, as well as "parents". This can be explained by the fact that many anti-vaccine proponents express their concerns regarding the vaccination of their children, claiming that vaccines have severe consequences for them. It is worth mentioning that "autism", which is another frequent word in this subcategory, is often perceived by some parents as a side effect of vaccines, even though there is no scientific evidence supporting such a connection.

Moreover, we observe that the vaccines "Covaxin" and "Novavax" are mentioned more frequently in the context of "anti_covid_vaccines" compared to "ant_all_vaccines", whereas "Pfizer" and "Moderna" appear more often in the context of "anti_all_vaccines".

Finally, it was predictable based on the temporal graphs that the term "flu shot" is mentioned much more frequently in the "anti_all_vaccines" group. However, it is intriguing to observe a higher frequency of the term "smallpox" in the "anti_covid_vaccines" category compared to the "anti_all_vaccines" category. One would anticipate a similar pattern to what we observe with the term "HPV vaccine", which refers to the vaccine for human papillomavirus, and it is more commonly brought up in the "anti_all_vaccines" category, which aligns with another prominent topic discussed in this category, namely pediatric vaccination and parental choice. This connection is especially relevant because the HPV vaccine is specifically recommended for administration in preteens at ages 11 or 12 years, with the vaccination series also able to be initiated as early as age 9 years [39].

To provide a concise comparison between anti-vaccine and pro-vaccine sentiments, we utilize the fourth category: "anti_vaccines", which encompasses uncategorized anti-vaccine tweets. In Figures 4.10 and 4.11, we observe the expected pattern: negative and conspiratorial words predominantly appear in the "anti_vaccines" category, while words encouraging vaccination are prevalent in the "pro_vaccine" category.

One notable term that arises is "big pharma", which reflects the conspiracy surrounding fears and suspicions toward pharmaceutical companies [40]. Another theory reflected in the data is the use of the bigram "Bill Gates", often associated with false claims about vaccines being used as a means to implant nanotechnology to control humanity [41]. Notably, "gene

therapy" appears with higher frequency in anti-vaccine tweets, despite it being one of the keywords used to gather tweets for all categories.

Lastly, while the term "herd immunity" is commonly mentioned across all categories, it is more prevalent in pro-vaccine conversations. Interestingly, the term "antivaxxer" is also more frequently utilized by pro-vaccine proponents.

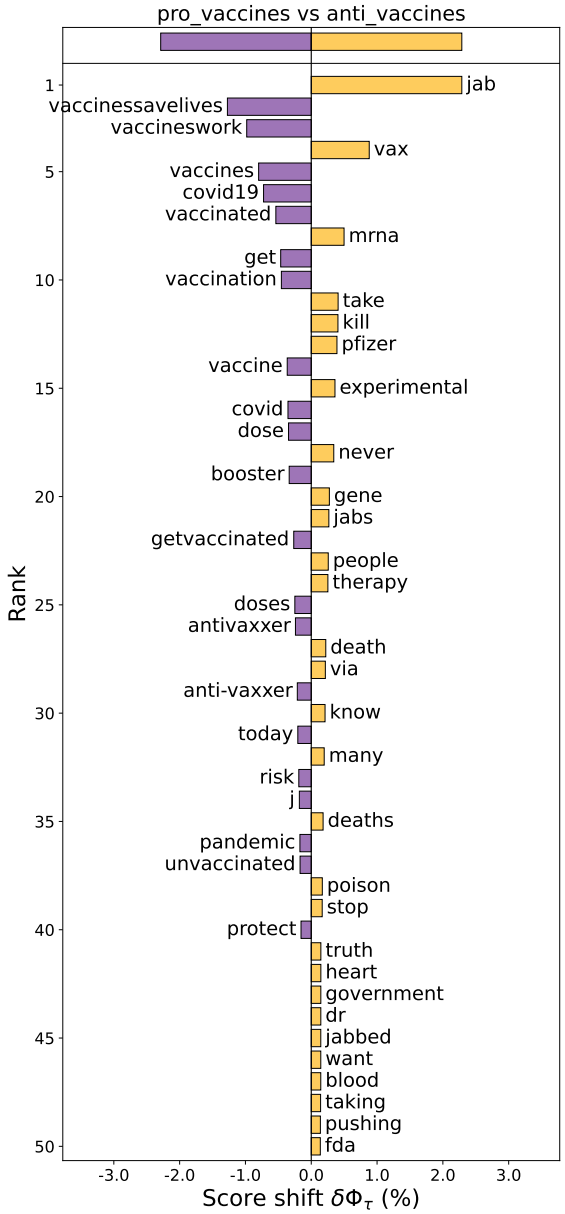

Figure 4.10: Shift: PV vs AV unigrams.

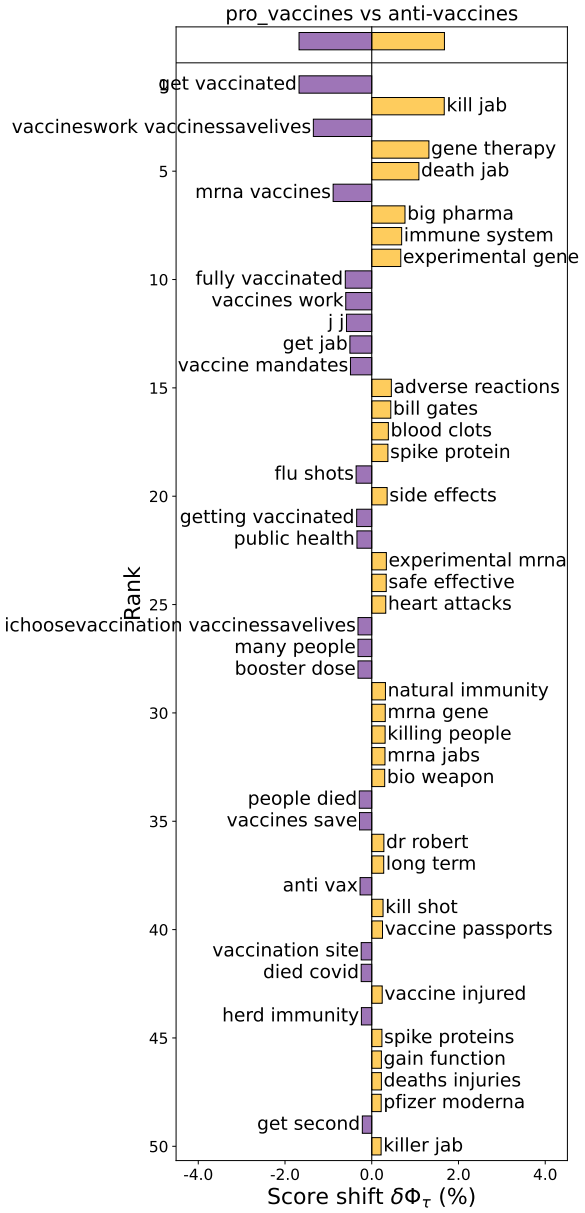

Figure 4.11: Shift: PV vs AV bigrams.

4.3.2 Scattertext

To enhance the visualization and analysis of language differences between categories, the Scattertext library in Python is utilized. Scattertext offers a visually appealing approach to explore and analyze text data by employing a scatter plot representation of terms in a two-dimensional space. Each point in the scatter plot corresponds to a specific term, with its position determined based on its relevance and frequency across different categories.

Scattertext provides a search feature that allows you to find words in their contextual messages. For that reason, we decided to keep the texts as intact as possible, performing minimal preprocessing by only removing users and URLs. To address potential challenges related to clutter and the presence of stopwords, two strategies are employed. Firstly, the minimum_term_frequency parameter is set to 20. This strategy aims to achieve a less cluttered graph by requiring a certain level of frequency for terms to be considered. Secondly, the term_significance parameter is assigned the st.LogOddsRatioUninformativeDirichletPrior() function. This function calculates the logarithm of the odds ratio, which compares a term's probability in a specific category with its overall probability across all categories. The use of a Dirichlet Prior helps prevent the undue influence of rare terms. As a result, stopwords are typically positioned in less prominent locations, allowing the focus to shift towards more significant and distinctive terms. This facilitates the identification of relevant patterns and features for analysis.

In Figure 4.12, we compare the vocabulary of the two categories of anti-vaccine proponents. At first glance, we can observe that this graph is much flatter and less scattered compared to the one that compares anti-vaccine proponents with pro-vaccine proponents. This indicates that these two subcategories have a much more similar vocabulary. This graph offers additional information compared to the previous one, as it allows us to search for any word and view the messages in which it appears, along with its corresponding frequency proportion.

To begin the comparison, let's focus on the terms that are closer to the upper-left quadrant, corresponding to words strongly associated with "anti_all_vaccines" and barely or not at all with "anti_covid_vaccines". Among the most polarizing terms, we find "autism" with a frequency proportion of 7:0. As mentioned in the introduction, this stems from the association between the MMR vaccine and autism, a belief that has extended to other vaccines. However, the fact that the frequency proportion is 7 to 0 suggests that it is not as prevalent in the context of COVID-19 vaccines. Another word related to the history of vaccines is "mercury", from the proportions, we can see that it is also not a concern within the "anti_covid_vaccines" group.

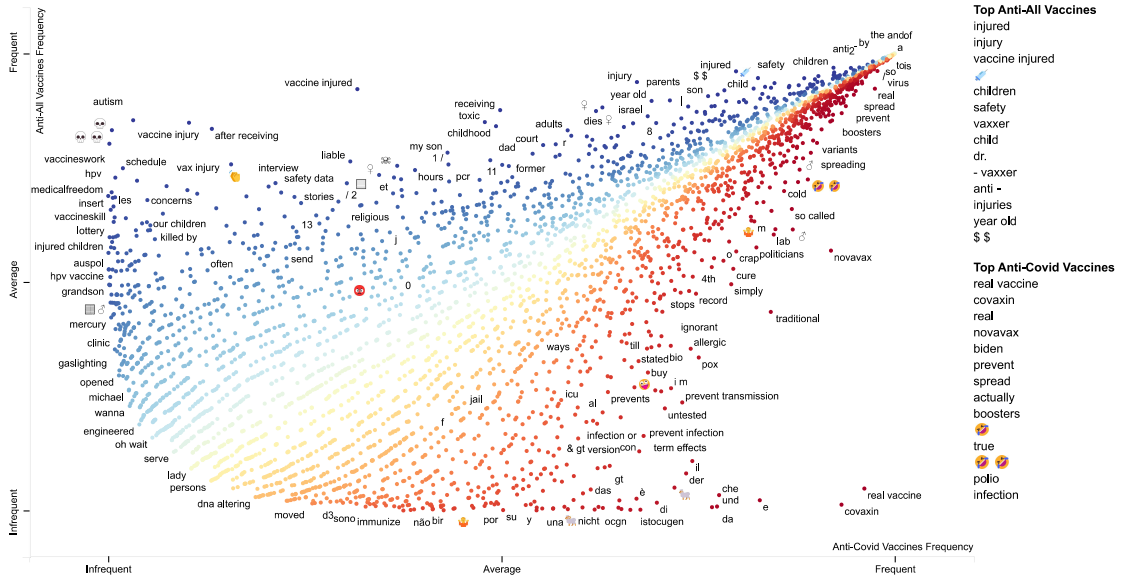

Figure 4.12: Comparison of anti_covid_vaccines and anti_all_vaccines using Scattertext.

Several hashtags were also found in the analysis. Firstly, a pro-vaccine hashtag, "#vaccineswork", which will be further analyzed. We also came across a new hashtag, "#medicalfreedom". The principles of the movement involve a strong dislike for government intervention in personal or family healthcare decisions [42], often accompanied by the promotion of extraordinary or miraculous remedies. As this is a long-standing movement, it makes sense that it is more trending among traditional vaccine skeptics. In general, we can observe that older conspiracies tend to appear more frequently in the "anti_all_vaccines" category.

Moving towards the lower-right quadrant, we find mentions of the Covaxin vaccine and Ocugen, which is the compary that holds the distribution rights for Covaxin in the US, and to a lesser extent, Novavax. This could be attributed to the fact that these two COVID-19 vaccines resemble traditional vaccines, specifically using attenuated virus and protein subunits, which could generate a slightly more positive sentiment among individuals in the "anti_covid_vaccines" group. However, this will be further analyzed in other sections. The "anti_all_vaccines" proponents do not seem to show significant interest in these two vaccines. However, we can observe that the COVID-19 vaccines from Pfizer, Moderna, and AstraZeneca, which have received considerable attention, have higher frequencies among the "anti_all_vaccines" proponents. We also notice that political terms like "politician" or "Biden" have higher frequencies among the "anti_covid_vaccines" group.

Additionally, while words related to childhood are characteristic of "anti_all_vaccines" proponents, the frequency proportion of the word "children" specifically is 37 to 13. This

implies a consistent number of tweets discussing the relationship between vaccines and childhood among the "anti_covid_vaccine" proponents as well.

In Figure 4.13, we can find the Scattertext graph that contrasts "anti_vaccines" with "pro_vaccines". Notably, for pro-vaccine tweets, a striking feature is the variety of hashtags used. Almost all the words found on the lower right quadrant, which have a close-to-zero frequency for anti-vaccine tweets, are hashtags in support of vaccines. In the previous analysis, we could already observe how "anti-vaxxer" is a term more commonly used by vaccine supporters. However, in this graph, we encounter a new term to criticize individuals with anti-vaccine beliefs, specifically those who are against COVID vaccines: "covidiots".

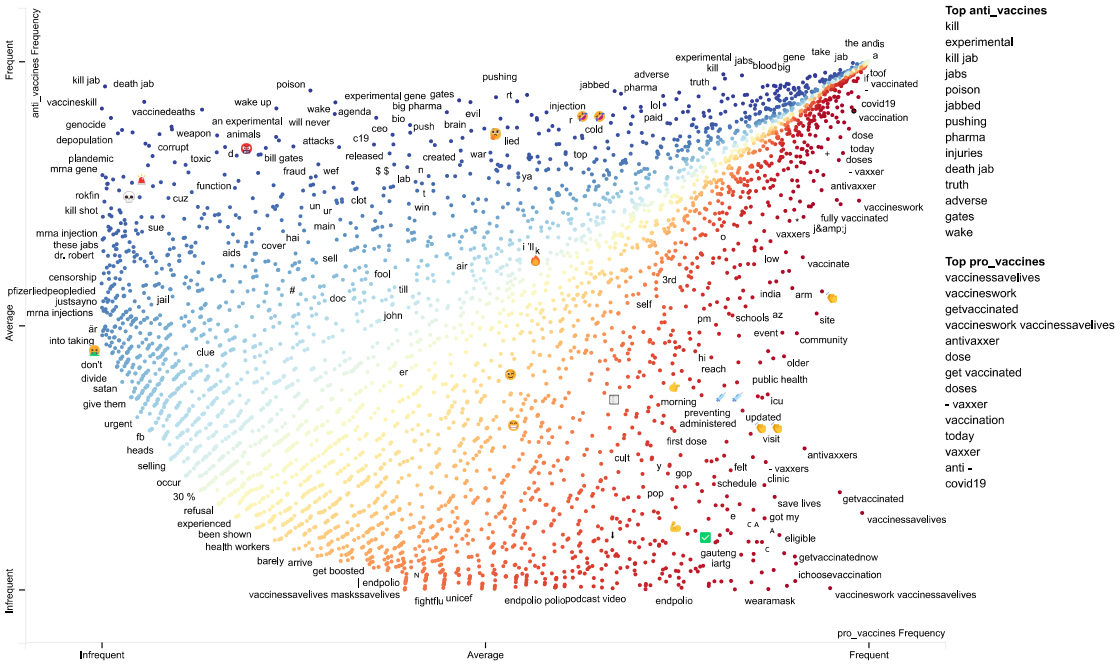

Figure 4.13: Comparison of Pro-Vaccines and Anti-Vaccines using Scattertext.

In the upper left quadrant, there are several conspiracy-related terms worth mentioning in relation to anti-vaccine tweets: "plandemic", "genocide", "depopulation" and "bio weapon". The term "plandemic" is a play on words combining "pandemic" and "plan" and suggests the belief held by some individuals that the pandemic was orchestrated. On the other hand, the other three terms imply the belief that vaccines are being used as a "weapon" to cause mass killings or to reduce the population. These terms highlight the deep mistrust in authorities and institutions among the anti-vaccine community, which could explain the frequent terms related to government, such as "corruption". The mention of political figures like "Biden" also indicates a focus on attributing responsibility to specific individuals within the government.

4.4 Sentiment analysis

For sentiment analysis, we employed Twitter-RoBERTa-base model, which has been fine-tuned on extensive Twitter datasets specifically for the detection of sentiment. In order to use this model, a simpler preprocessing step was performed, which involved removing URLs, usernames and tokenization. Then, each tweet is processed by the model to obtain scores for various sentiment classes. These scores are then normalized using the softmax function. The sentiment class with the highest score is identified, and the corresponding sentiment label ("NEGATIVE", "NEUTRAL" or "POSITIVE") is assigned to each tweet based on the highest score obtained.

In Figure 4.14, we can see that the predominant sentiment for all categories is negative. In fact, the three types of anti-vaccine categories have very similar percentages of neutrality, positivity, and negativity. On the other hand, in the "pro_vaccines" group, the percentage of neutrality is similar to the percentage of negativity, and we can observe approximately four times more positivity compared to the anti-vaccine categories.

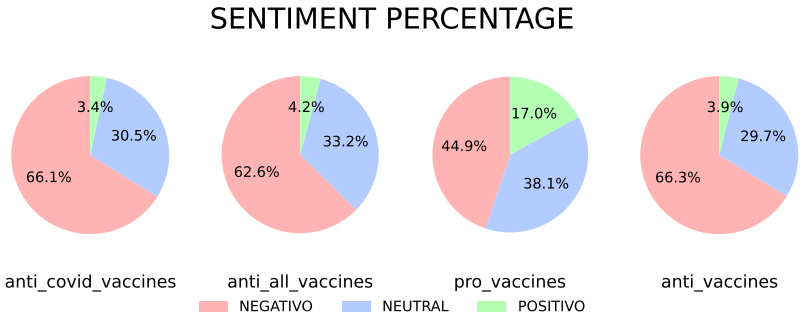

Figure 4.14: Percentage of Sentiment by Category.

To have a general overview of the word distribution among categories, a wordcloud is generated for each sentiment within each category. It can be observed that the word-clouds do not provide much information as they repeat and predominantly contain the same words across all categories (Fig. 4.15). Hence, we opted for a more simplified visualization approach.

To achieve this, the top 100 most frequent bigrams and unigrams were obtained for each category, and those that intersected in at least two of the three categories were selected. Additionally, some n-grams that did not contribute any new information were removed. However, in the case of unigrams, "Novavax" and "Covaxin" were included, even though they did not meet the previous condition. This decision was based on the insights gained from the previous analyses, indicating that analyzing the sentiment associated with these

terms would be interesting.

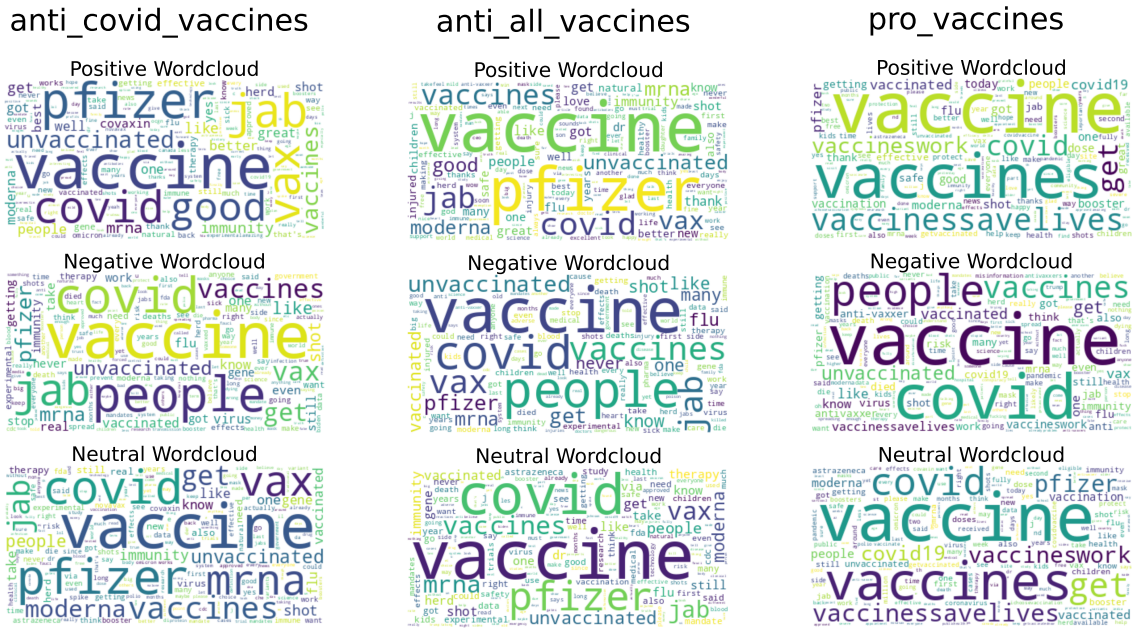

Figure 4.15: Wordclouds of sentiment analysis for different categories: a) anti_covid_vaccines, b) anti_all_vaccines and c) pro_vaccines.

In Figures 4.16 and 4.17, one notable observation is the higher percentage of neutrality reflected in the pro-vaccine category compared to the other categories. Specifically, AstraZeneca, Moderna, Pfizer, Covaxin, and Novavax are mostly mentioned in a neutral context. When examining the bigrams, we find two vaccines that utilize viral vector technology, Johnson & Johnson and the Sputnik V Russian vaccine, both for COVID-19. For pro-vaccines, they also have more prevalence of neutrality but we observe less positivity for this vaccines compared to the other vaccines.

It is interesting to note, however, that the term "vax" appears in a significantly higher percentage of negative tweets compared to the synonymous terms "jab" and "vaccine". The term "booster", which was prominent in the temporal graphs for the "pro_vaccines" category, maintains a positive and neutral sentiment, indicating that booster doses were well-received among vaccine supporters.

Turning our attention to the two anti-vaccine categories, we observe a similar pattern. The majority of terms in both categories tend to lean towards negativity, prompting us to explore the percentages of positivity and neutrality. In the case of "anti_all_vaccines", most terms show a higher proportion of neutrality than in the "anti_covid_vaccines" category. This suggests that individuals in this category may sometimes express their beliefs in a more informative or explanatory manner, rather than consistently exhibiting strong negativity,

even if their perspectives are rooted in misinformation or misinterpretation. Notably, for the Moderna vaccine, the highest count in this category is associated with a neutral sentiment.

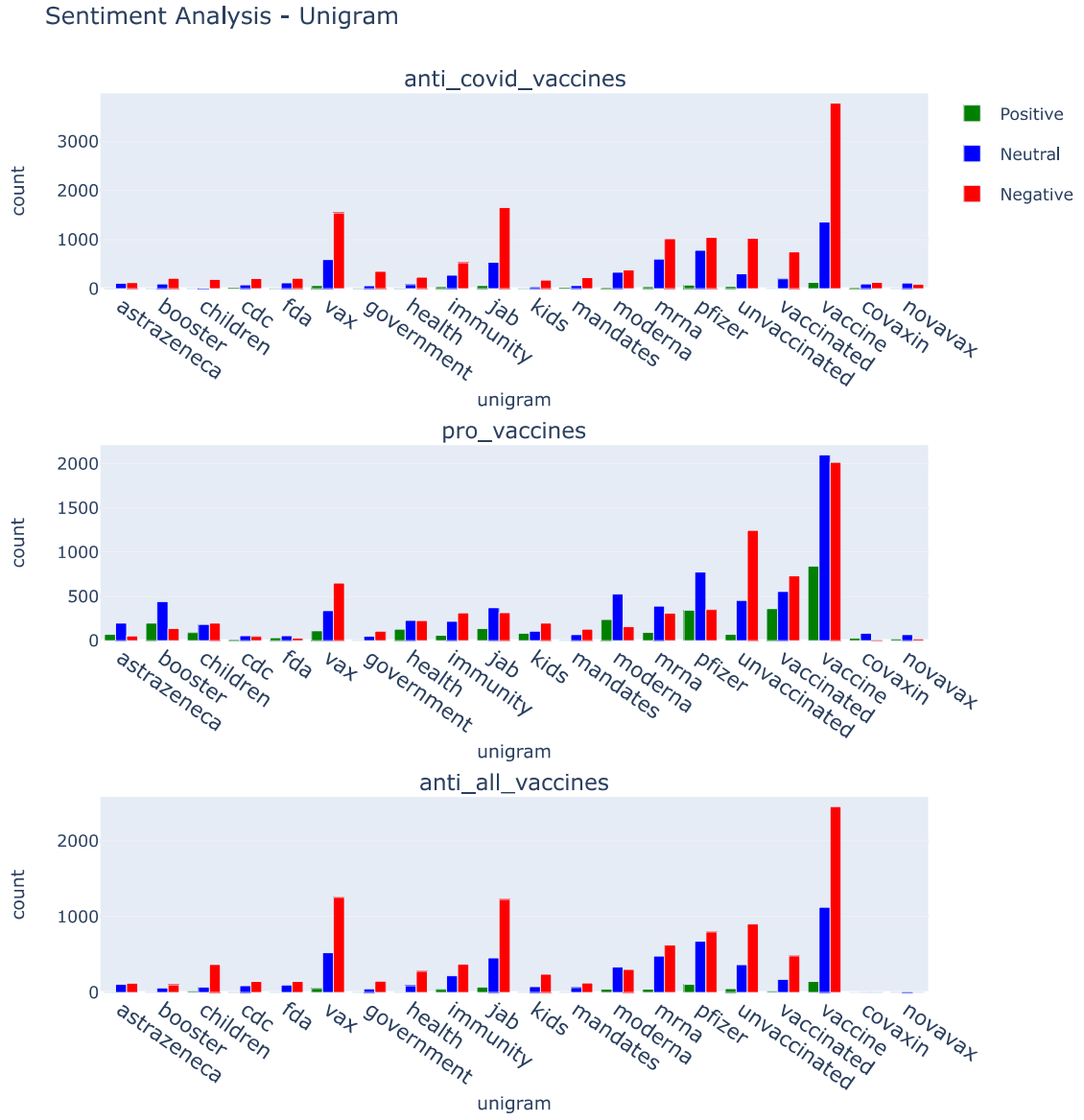

Figure 4.16: Distribution of sentiment scores for individual words in the dataset.

Among the "anti_covid_vaccines" group, Novavax is the only word for which the percentage of neutrality (approximately 54%) surpasses that of negativity, while Covaxin has the highest percentage of positivity, at 11.2%. This is significant considering that the next word with the highest positivity percentage after Covaxin is "immunity" with 4.92% positivity. It is worth noting that Novavax is mentioned only 25 times and Covaxin only 4 times within the "anti_all_vaccines" group, suggesting a lack of specific interest in these two vaccines (Fig. 4.16).

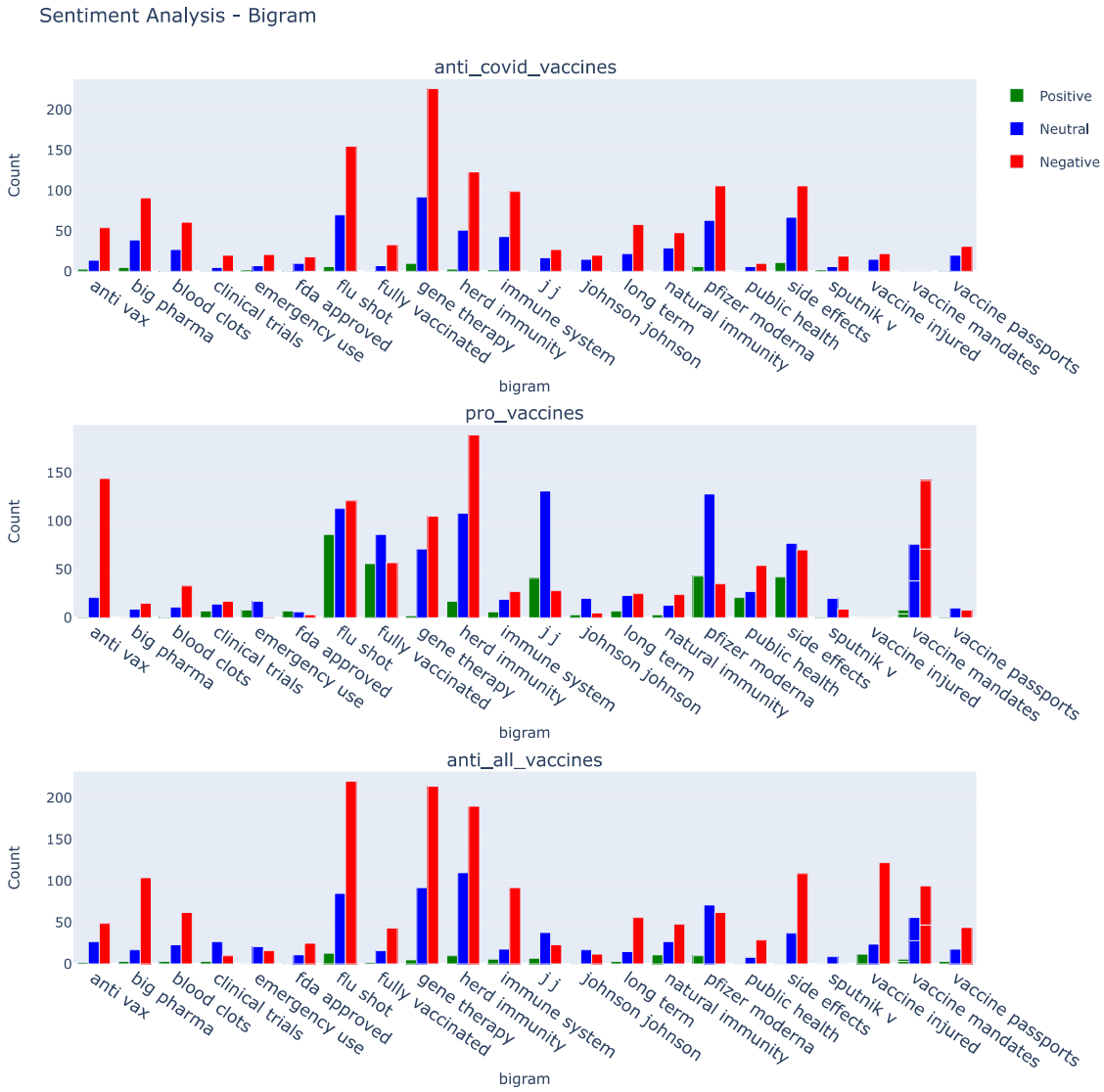

Figure 4.17: Distribution of sentiment scores for word pairs (bigrams) in the dataset.

It is evident that "flu shot" predominantly appears in a negative context across all three categories (Fig. 4.17). The highest percentage of negativity is observed in the "anti_all_vaccines" category, while the highest percentage of positivity is associated with the pro-vaccine individuals. The flu shots have gained attention as they share similar criticisms with COVID vaccines. Both vaccines target specific strains, necessitating seasonal flu shots and boosters for emerging COVID variants. Furthermore, there is no guarantee of complete protection against infection, although they effectively decrease the risk of severe complications [43]. This observation could help explain the negative sentiment surrounding flu shots.

Regarding regulations, the term "emergency use" refers to Emergency Use Authorization

(EUA), a regulatory mechanism that grants accelerated access to medical countermeasures during public health emergencies like the COVID-19 pandemic [44]. It allows the FDA to authorize the use of medical products, including vaccines. The positivity for this mechanism is higher in the pro-vaccines category and more negative in the anti-vaccine categories. Other terms related to vaccine regulations, such as "mandates", "vaccine mandates" and "vaccine passports", are predominantly mentioned in a negative context across all three categories. However, there is a slightly more neutral and positive sentiments observed in among pro-vaccines individuals compared to the anti-vaccine categories.

4.5 Hashtag analysis

After examining the distribution of hashtags, we can see that there is a significantly higher occurrence of hashtags in the pro-vaccine category. Table 4.1 shows the number of hashtags in each category.

Category	**Number of Hashtags**
anti_covid_vaccines	742
anti_all_vaccines	1,148
pro_vaccines	3,182

Table 4.1: Number of hashtags per category.

One notable observation is the significant presence of the hashtag "#covaxin" in the wordcloud associated with "anti_covid_vaccines" (Fig. 4.18). Additionally, there are several hashtags containing the word "Covaxin", such as "#wechoosecovaxin", "#ichoosecovaxin", and "#covaxin4kids". However, it is important to note that these hashtags have relatively low counts. Interestingly, these hashtags have no occurrences in the pro-vaccines category, despite the fact that "Covaxin" was one of the keywords used to gather tweets from users in all four categories. This indicates that individuals in the "anti_covid_vaccines" group are not using these hashtags to mock vaccine supporters, as these hashtags are not even being used in the "pro_vaccines" category. Instead, it suggests that there is a minority that could support these vaccines due to the fact that they are the only ones that utilize the same technologies as traditional vaccines. The same pattern may apply to Novavax, which was the only word that appeared to be associated with a neutral sentiment for this category in the previous analysis (Fig 4.16), but it is not as conclusive as with Covaxin.

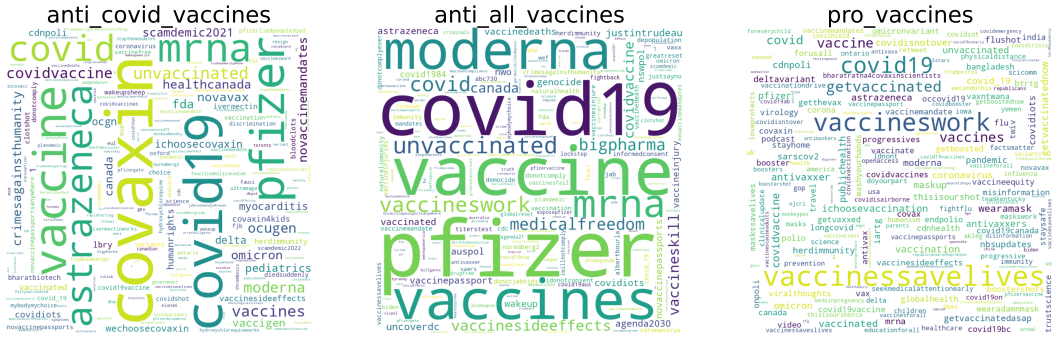

Figure 4.18: Hashtag Wordclouds per category.

Additionally, it is worth noting that within the "anti_all_vaccines" category, there are a few mentions of hashtags such as "#vaccineswork" and "#vaccinessalives". While the hashtag "#vaccinessavelives" has a minimal count in this category, "#vaccineswork" appears 46 times. Since these are two of the most frequently used hashtags among pro-vaccine individuals, this time, there are greater chances of anti-vaccine proponents using these hashtags in a sarcastic manner to ridicule vaccine advocates.

For that reason, an analysis of irony was conducted utilizing the "roBERTa-base" model which is fine-tuned for irony detection. Specifically, the percentage of irony was examined for the hashtags "#vaccinessavelives" and "#vaccineswork" within the anti-all vaccines category.

From the irony analysis (Fig. 4.19), it was determined that 33.33% of the mentions of "#vaccinessavelives" and 26.67% of the occurrences of "#vaccineswork" in the "anti_all_vaccines" category were used ironically. In contrast, the ironic mentions within the pro-vaccines category were 0.76% and 1.1% for the respective hashtags.

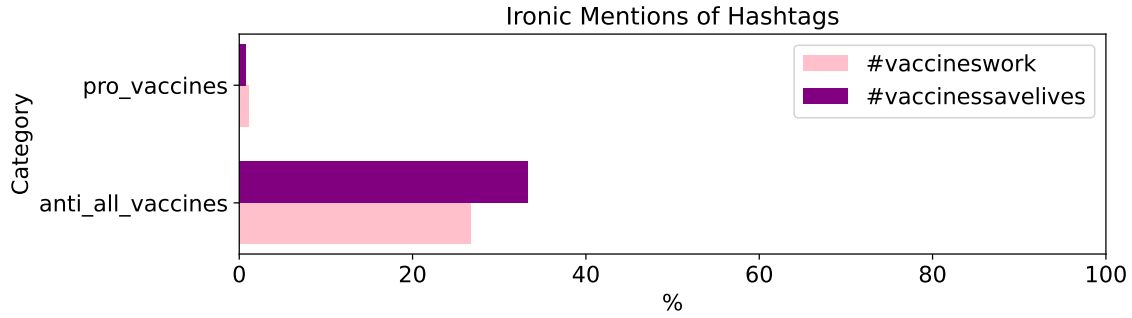

Figure 4.19: Percentage of ironic mentions for #vaccineswork and #vaccinessavelives.

CHAPTER 5

Classifier Evaluation

5.1 Introduction

In this chapter, two classifiers, Linear Support Vector Classification and Random Forest, will be employed to attempt categorizing the dataset into the "anti_all_vaccines", "pro_vaccines" and "anti_covid_vaccines" categories. The "anti_vaccines" category is not considered for classification purposes, since it comprises uncategorized tweets that may include users who hold "anti_all_vaccines" or "anti_covid_vaccines" viewpoints. The objective is to evaluate the accuracy of the classifiers in categorizing the tweets and determine if there is a significant difference in the language used between the three remaining categories.

5.2 Linear Support Vector Classification

For this analysis, the text preprocessing involved spell correction, tokenization of tweets, URL removal, elimination of stopwords, and punctuation. Next, the dataset was divided into training and test sets, with the test set comprising 33% of the total data. This division allows for evaluating the performance and generalization capabilities of the classifiers on unseen data. Additionally, text vectorization was performed with a maximum limit of 2000

features to transform the textual data into numerical representations suitable for machine learning algorithms.

First, Linear Support Vector Classification (Linear SVC) algorithm was assessed. It aims to find a hyperplane that maximizes the distance between classified samples. Table 5.1 illustrates the results obtained for Linear SVC.

Category	**Precision**	**Recall**	**F1-Score**	**Support**
anti_all_vaccines	0.51	0.50	0.50	3889
anti_covid_vaccines	0.56	0.58	0.57	4454
pro_vaccines	0.71	0.71	0.71	4982
Accuracy	-	-	0.60	13325
Macro Avg	0.60	0.60	0.60	13325
Weighted Avg	0.60	0.60	0.60	13325

Table 5.1: Linear SVC Classification Results.

We can observe that the "pro_vaccines" category has a significantly higher F1-score compared to the other two categories, with a score of around 71% while the rest are around 50%. This suggest indicate that the "pro_vaccines" category, being the easiest to classify, may indeed have a more distinctive vocabulary compared to the other categories. However, to further investigate which categories tend to be more frequently misclassified or confused with one another, we will analyze the confusion matrix (Table 5.2).

	anti_all_vaccines	**anti_covid_vaccines**	**pro_vaccines**
anti_all_vaccines	1928	1256	705
anti_covid_vaccines	1170	2570	714
pro_vaccines	650	774	3558

Table 5.2: Confusion Matrix of Linear SVC.

In the (0,1) position of the confusion matrix, we observe 1256 instances misclassified as "anti_covid_vaccines" when they were actually tweets belonging to the "anti_all_vaccines" category. Conversely, in the (1,0) position, we have 1170 tweets from the "anti_covid_vaccines"

category classified as “anti_all_vaccines”. This indicates that the most significant confusion occurs between these two categories.

Interestingly, among the misclassified “pro_vaccines” tweets, a higher degree of confusion is observed with the “anti_covid_vaccines” category compared to ”anti_all_vaccines”. This indicates a stronger linguistic similarity or association between the language used in “pro vaccines” and “anti_covid_vaccines”, in contrast to the other anti-vaccine category.

5.3 Random forest

Random forest is a machine learning algorithm that combines the outputs of multiple decision trees to make predictions. The algorithm works by creating an ensemble of decision trees, each trained on a different subset of the data. The predictions from the individual trees are then combined to make a final. The results and confusion matrix are presented in Tables 5.3 and 5.4, respectively.

Category	**Precision**	**Recall**	**F1-Score**	**Support**
anti_all_vaccines	0.56	0.43	0.49	3889
anti_covid_vaccines	0.55	0.62	0.59	4454
pro_vaccines	0.69	0.73	0.71	4982
Accuracy	-	-	0.61	13325
Macro Avg	0.60	0.60	0.60	13325
Weighted Avg	0.61	0.61	0.60	13325

Table 5.3: Random Forest Classification Results.

	anti_all_vaccines	**anti_covid_vaccines**	**pro_vaccines**
anti_all_vaccines	1691	1396	802
anti_covid_vaccines	862	2767	825
pro_vaccines	482	840	3660

Table 5.4: Confusion Matrix of Random Forest.

With Random Forest, we obtain similar results, although slightly worse overall (Table 5.3). In Table 5.4, we can see that Random Forest tends to correctly label more tweets as "anti_covid_vaccines", but it also has a higher number of false positives for this category. On the other hand, it demonstrates fewer errors in mislabeling tweets as "anti_all_vaccines", indicating a tendency to over-label anti-vaccine tweets as "anti_covid_vaccines".

For the "pro_vaccines" category, the results are similar to the previous algorithm. However, it becomes more evident that tweets classified as "pro_vaccines" have more linguistic similarities with "anti_covid_vaccines" than with "anti_all_vaccines".

CHAPTER 6

Conclusions and future work

In this chapter, we discuss the conclusions drawn from this project, the accomplished objectives, the challenges encountered, and suggestions for future work.

6.1 Conclusions

First and foremost, regarding the dataset capture, it has been observed that it is quite easy to find various types of conspiratorial messages on an open social network like Twitter. Moreover, on social media platforms, where we seek quick and summarized information, we often do not dedicate enough time to verify the sources of what we read. This can lead to the spread of false or misinterpreted information, especially within conspiratorial communities that tend to reinforce each other.

In the case of anti-vaccine proponents, a broad spectrum of individuals with different beliefs and motivations has been identified. Some individuals may have legitimate concerns about the newly emerged vaccines and choose to wait for more information before deciding to get vaccinated, while others may hold more extreme conspiratorial theories, such as claiming that vaccines are government inventions to control the population. This diversity of beliefs and motivations makes it challenging to categorize all anti-vaccine proponents under

a single label. For that reason, classifying them into traditional anti-vaccine proponents and those specifically against COVID-19 vaccines can be useful in better understanding their distinct perspectives. However, after the analysis, we can conclude that in recent years, both categories have been using very similar language. Therefore, there may be many other ways to classify them in order to present more distinct differences between them.

Regarding the data analysis, the "anti_all_vaccines" category, which refers to users discussing a wide range of vaccines before the existence of COVID-19, exhibited a more focused approach in discussing COVID-19 and its vaccines during recent years. This reflects how the COVID-19 pandemic has profoundly impacted our lives and dominated conversations on social media for several months, overshadowing discussions about other vaccines. Additionally, it is evident that while pro-vaccine proponents and anti-vaccine proponents hold distinct and even opposing viewpoints, they both tweet about the same topics and are often driven by the same events. Consequently, there are many concrete terms that they share in common. Nevertheless, the sentiment analysis reveals significant differences in context, tone, and intention behind their language usage.

Lastly, in the classifier analysis, it is notable that the vocabulary between traditional anti-vaccine proponents and those solely against COVID-19 vaccines appears to be very similar, to the extent that algorithms struggle to differentiate between them. It can also be observed that the communication style of "anti_covid_vaccines" proponents on Twitter is more aligned with pro-vaccine proponents than with the other category of anti-vaccine users.

To conclude, it is worth noting that all these conclusions were obtained by demonstrating the immense utility of Natural Language Processing (NLP) in analyzing and understanding online discussions. NLP techniques have proven to be highly effective in extracting meaningful insights from large volumes of text data.

6.2 Achieved goals

G1 Gather tweets related to vaccines from both pro-vaccine and anti-vaccine individuals and store them in a database.

We successfully collected approximately 39,000 tweets from anti-vaccine individuals and around 15,000 tweets from pro-vaccine individuals. These tweets were stored in a database for further analysis.

G2 Perform a natural language processing (NLP) analysis on the collected

tweets to identify common patterns, themes, and trends:

NLP analysis was conducted to examine various aspects such as temporal analysis of terms, vocabulary differences between categories, sentiment analysis, and hashtag analysis. We identified common patterns and themes within the collected tweets, providing insights into the language used by individuals with different stances on vaccines. However, further analysis is required to gain a comprehensive understanding of anti-vaccine sentiments and related factors.

G3 Utilize a classification model, to automatically distinguish between the categories of the database, aiming to gain a better understanding of the language and distinctive features of each stance:

We implemented two classification algorithms, Random Forest and Linear SVC, to classify the tweets into pro-vaccine and anti-vaccine categories. The results of these algorithms provided meaningful feedback, highlighting the differences in vocabulary and language usage between the two categories.

6.3 Challenges encountered

During the analysis process, we encountered several issues that affected both the process itself and the obtained results. Firstly, when searching for tweets from anti-vaccine users categorized as "anti_all_vaccines" in the past and subsequently retrieving tweets from these users over the past three years, a significant number of users were lost. As a result, this category had the lowest number of tweets available for analysis. Furthermore, recent changes in Twitter's policy required modifications to the tools used for scraping tweets, making it more challenging to extract tweets from the most recent months [45]. Additionally, categorizing tweets without proper context proved to be a difficulty encountered during the analysis.

During the data analysis process, certain terms emerged from users who excessively tweeted about the same topic. To ensure data quality, these users were removed from the dataset. As a result, the process was not entirely linear and required iterative refinement.

Creating informative graphs posed a challenge due to the large volume of information. This was addressed by utilizing interactive visualizations, allowing for better organization and presentation of the data. However, when including static versions of these interactive visualizations, some interactive features and details may be lost, potentially reducing the amount of information communicated.

6.4 Future work

This section outlines the potential future lines of research for this project, aiming to expand the understanding of the dynamics and drivers behind anti-vaccine discourse on social media. To achieve a comprehensive understanding, future research could involve comparing different social media platforms such as Facebook and Reddit. Analyzing these platforms can provide insights into the variations in content, engagement, and discourse related to vaccine hesitancy.

Furthermore, it could be interesting to expand the investigation by incorporating user demographic information. Factors such as age, location, educational background, and socioeconomic status may influence the prevalence and intensity of anti-vaccine sentiments.

Additionally, analyzing the vocabulary of vaccine skeptics on social media both before and after the COVID-19 pandemic can reveal changes in language use. By exploring these linguistic variations, researchers can gain valuable information about vaccine hesitancy across a wider range of vaccines.

Anti-vaccine sentiments often intersect with conspiratorial beliefs. Conducting a more focused analysis on the relationship between conspiratorial beliefs and anti-vaccine discourse could reveal factors driving vaccine hesitancy. Also, understanding the political ideologies associated with different anti-vaccine categories can provide valuable observations into the social and cultural factors influencing anti-vaccine movement.

Lastly, future research could explore and evaluate additional classifier algorithms to enhance the accuracy and effectiveness of NLP models in detecting and categorizing anti-vaccine content. It would contribute to optimizing the identification and classification of anti-vaccine sentiment.

APPENDIX A

Impact of this project

This appendix reflects, quantitatively or qualitatively, on the possible impact on society, economics, environment and ethics.

A.1 Social impact

By analyzing data related to anti-vaccine discourse, healthcare systems will be able to develop more impactful vaccination campaigns that utilize appropriate communication channels, clear and understandable messages, and address specific fears and doubts that individuals may have. This approach aims to encourage the population to trust vaccines, reduce vaccine hesitancy, and promote the dissemination of accurate and verified vaccine information. Consequently, vaccination rates will significantly increase, contributing to the establishment of a society that recognizes the value and benefits of vaccines.

Through these efforts, the project aims to prevent the spread of diseases and establish a safer and healthier environment for the entire community. This includes protecting the most vulnerable groups who may be at risk of severe illnesses or complications, such as children, the elderly, and individuals with weakened immune systems. Furthermore, vaccination contributes to the protection of vulnerable groups such as children, the elderly, and

individuals with weakened immune systems.

A.2 Economic impact

By tackling vaccine hesitancy and consequently increasing vaccination rates, the project yields significant economic benefits. Firstly, it leads to a healthier and more productive workforce. By reducing vaccine-preventable diseases, the need for costly medical treatments and hospitalizations is diminished. This not only saves healthcare expenses but also relieves the burden on healthcare systems, allowing for more efficient allocation of resources.

Moreover, vaccinated individuals have a lower risk of falling ill, resulting in decreased absenteeism and sustained levels of productivity among workers. This helps businesses maintain their operations smoothly and prevents disruptions in economic activities.

Furthermore, certain sectors heavily reliant on public interaction, such as tourism, hospitality, and entertainment, experienced substantial economic losses during the pandemic. However, through the establishment of a safer environment via widespread vaccination, these sectors can recover more swiftly. Not only does this facilitate the recovery from past pandemics, but it also builds resilience against potential future ones.

A.3 Enviromental impact

The pandemic has brought forth significant environmental challenges. These include the increase in non-recyclable waste, the generation of substantial quantities of organic waste due to diminished agricultural and fishery exports, and the difficulties in maintaining and monitoring natural ecosystems [46]. By comprehending and mitigating the anti-vaccine discourse, we can foster vaccine confidence and, ultimately, enhance immunization rates. This is crucial to expedite the recovery process and restore a safe environment for resuming normalcy in agricultural and fishery activities, as well as enabling ecosystem monitoring personnel to resume their vital roles. Furthermore, by preventing future pandemic disasters, we can effectively address the root causes of these environmental issues.

In addition to the aforementioned effects, by increasing vaccine acceptance, it is possible to mitigate the transmission of zoonotic diseases, which are those that can be transmitted between animals and humans. By vaccinating both animals and humans, we can effectively break the chain of infection and prevent the spillover of diseases from wildlife to humans or domesticated animals. This proactive approach not only safeguards public health but also

contributes to the preservation of ecological balance and the well-being of diverse species.

A.4 Ethical impact

The project demonstrates a strong commitment to the ethical principle of integrity by actively promoting the dissemination of accurate, evidence-based information about vaccines. By addressing and countering the misinformation and myths that surround vaccines, it fosters transparency, honesty, and the responsible use of data. This approach builds trust in the healthcare system and scientific community, which are crucial for the principles of autonomy and informed decision-making. It empowers individuals to make informed choices regarding their own health and well-being, based on accurate and reliable information. This promotes individual agency and respects their right to make choices aligned with their values and beliefs.

APPENDIX B

Economic budget

This appendix details an adequate budget to bring about the project.

B.1 Equipment

The physical resource utilized for this project was a laptop equipped with an Intel Core i7-1165G7 2.80 GHz x4 CPU, 16GB of RAM, and 512GB of SSD storage. The approximate cost of this laptop is **900 euros**.

B.2 Human resources

The total time invested in the project is equivalent to 12 ECTs, which corresponds to approximately 360 hours.

Considering that there is only one worker assigned to this project and an intern engineer receives a monthly salary of around 500 euros for working part-time, with a daily commitment of 4 hours and excluding weekends, we can estimate the duration of the project.

The 360 hours can be divided into 90 days, based on 4 hours of work per day. Taking

into account an average of 22 working days per month (considering a month of 30 days), the project duration amounts to approximately 4.09 months.

Multiplying the monthly salary of 500 euros by the duration of 4.09 months, we estimate the total human budget for the project to be around **2,045 euros**.

B.3 Sofware licenses

All the software programs used in this project were open source, which means that no expenses were incurred for licenses.

B.4 Total budget

The total budget for this project is around 2,945.00€.

Item	**Hours worked**	**Monthly salary**	**Subtotal**
Worker	360	500 €	2,045.00€
PC	-	-	900.00€
		Total	**2,945.00€**

Table B.1: Cost Breakdown

APPENDIX C

Supporting graphs

In this appendix, we present the second part of the vocabulary evolution graphs. We include both parts to ensure that the complete legend is displayed, as the interactivity was lost when static versions of the graphs had to be included in this document.

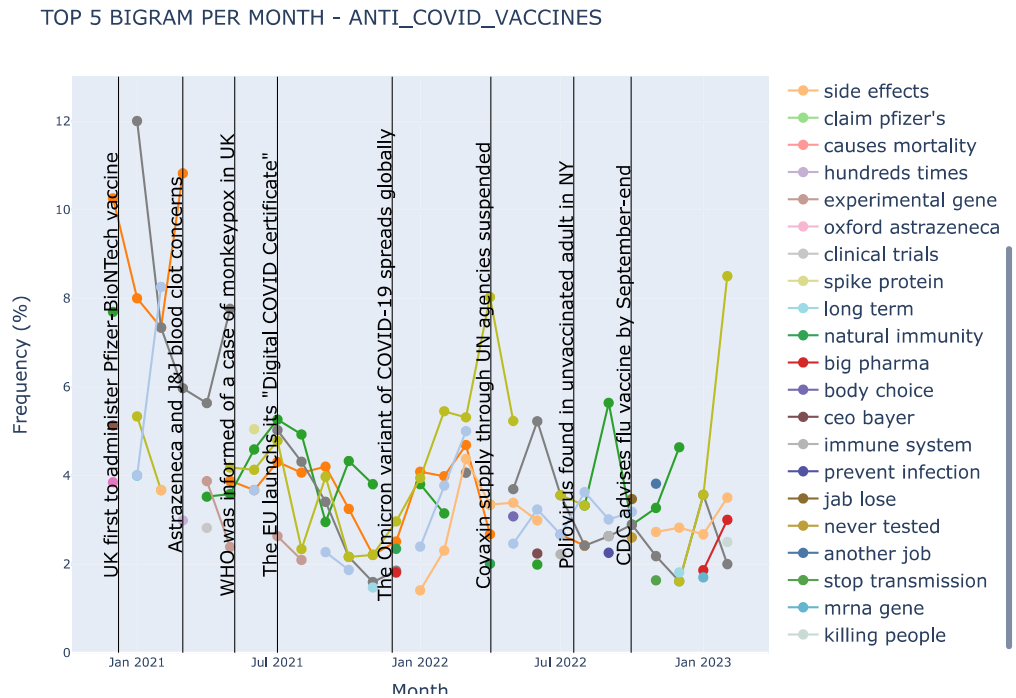

Figure C.1: Part II: Evolution of bigrams in the "anti_covid_vaccines" category.

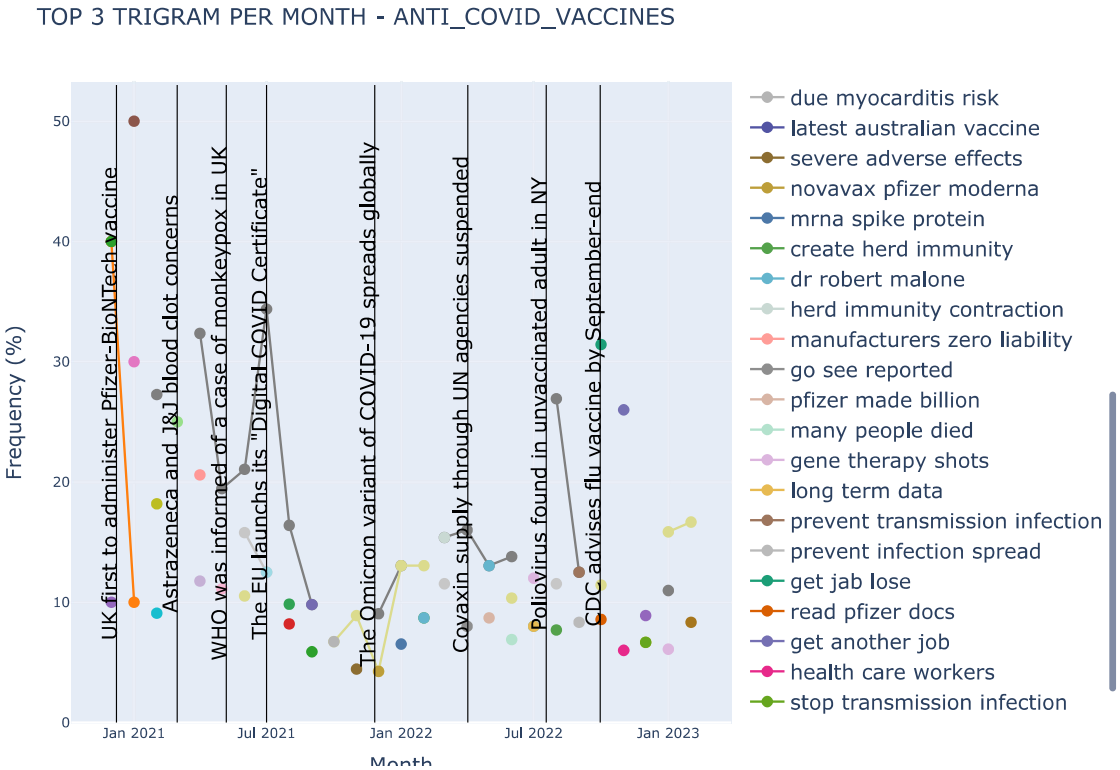

Figure C.2: Part II: Evolution of trigrams in the "anti_covid_vaccines" category.

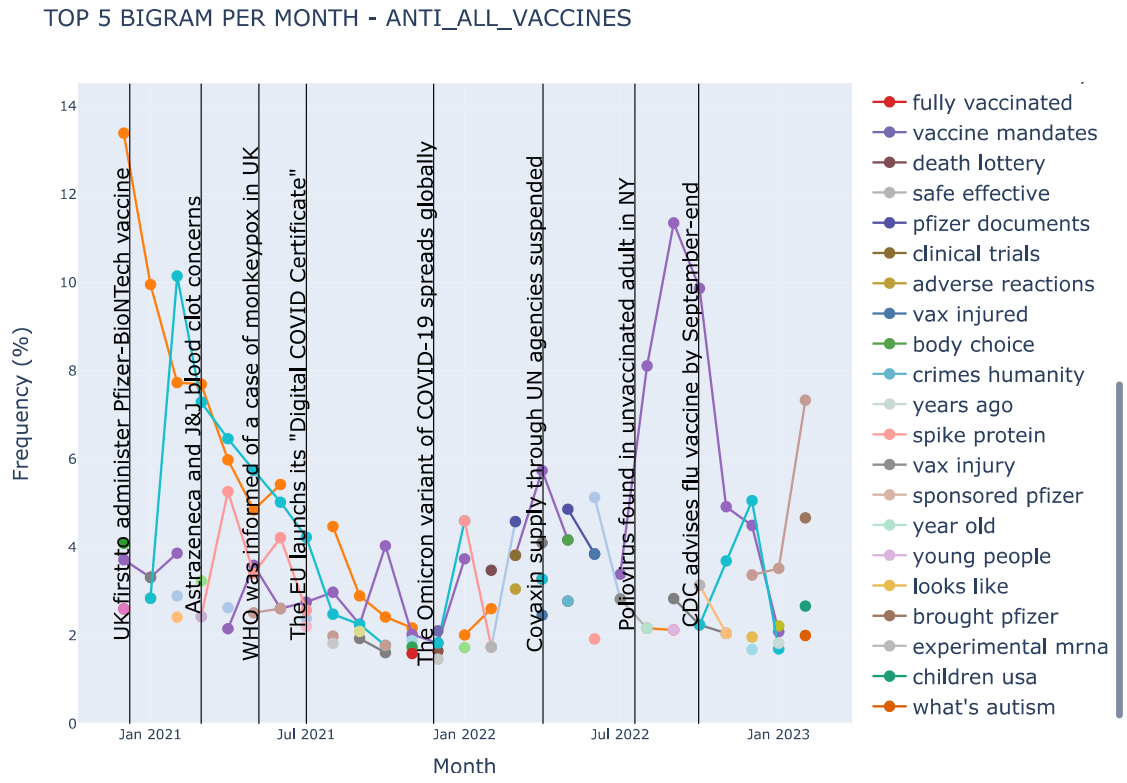

Figure C.3: Part II: Evolution of bigrams in the "anti_all_vaccines" category.

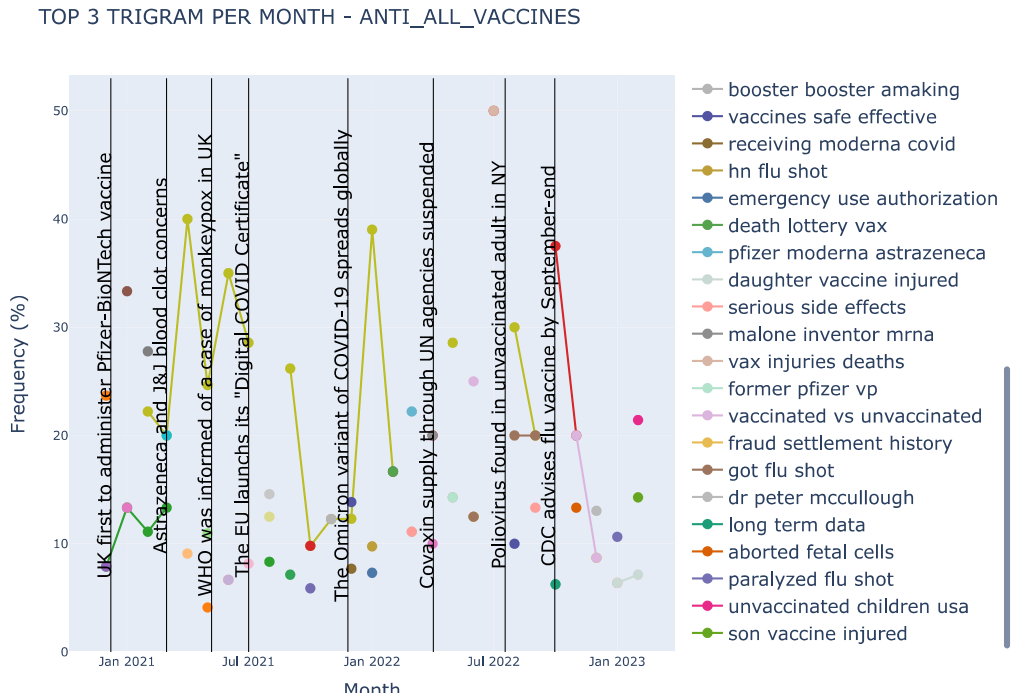

Figure C.4: Part II: Evolution of trigrams in the "anti_all_vaccines" category.

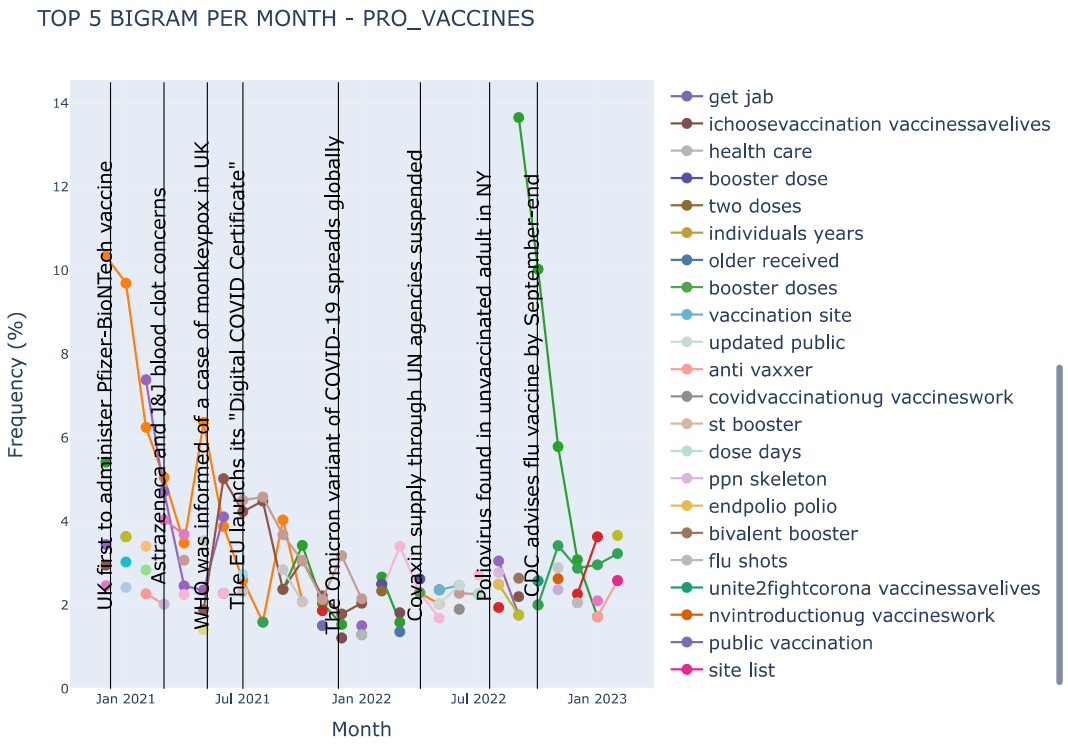

Figure C.5: Part II: Evolution of bigrams in the "pro_vaccines" category.

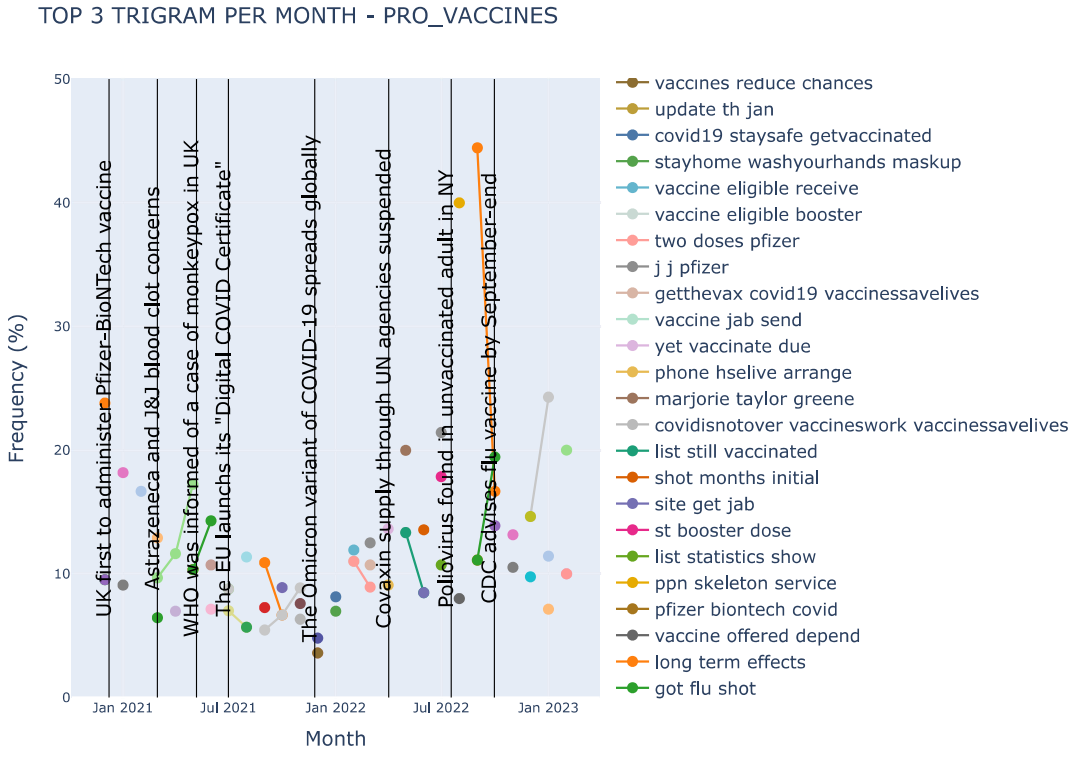

Figure C.6: Part II: Evolution of trigrams in the "pro_vaccines" category.

Bibliography

[1] R. F. Nuwarda, I. Ramzan, L. Weekes, and V. Kayser. Vaccine hesitancy: Contemporary issues and historical background. *Vaccines*, 10(10):1595, 2022.

[2] World Health Organization. Who coronavirus (covid-19) dashboard. Available online at https://covid19.who.int/. Accessed: 2023-05-20.

[3] World Health Organization. Vaccines and immunization. Available online at https://www.who.int/health-topics/vaccines-and-immunization. Accessed: 2023-06-11.

[4] U.S. Department of Health & Human Services. Vaccine types. Available online at https://www.hhs.gov/immunization/basics/types/index.html. Accessed: 2023-05-19.

[5] World Health Organization. Suspension of supply of covid-19 vaccine covaxin through un agencies by who. Available online at https://www.who.int/news/item/02-04-2022-suspension-of-supply-of-covid-19-vaccine-covaxin-through-un-agencies-by-WHO. Accessed: 2023-04-30.

[6] World Health Organization. The sinopharm covid-19 vaccine: what you need to know. Available online at https://www.who.int/news-room/feature-stories/detail/the-sinopharm-covid-19-vaccine-what-you-need-to-know, 2021. Accessed: 2023-06-12.

[7] Centers for Disease Control and Prevention. Understanding how covid-19 vaccines work. Available online at https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/how-they-work.html. Accessed: 2023-05-20.

[8] Pfizer. mrna technology: What it is and how it works. Available online at https://www.pfizer.com/science/innovation/mrna-technology. Accessed: 2023-05-19.

[9] Genomics Education Programme. Why mrna vaccines aren't gene therapies. Available online at https://www.genomicseducation.hee.nhs.uk/blog/why-mrna-vaccines-arent-gene-therapies/, 2021. Accessed: 2023-05-19.

[10] J. Romate, E. Rajkumar, A. Gopi, J. Abraham, J. Rages, R. Lakshmi, J. Jesline, and S. Bhogle. What contributes to covid-19 vaccine hesitancy? a systematic review of the psychological factors associated with covid-19 vaccine hesitancy. *Vaccines*, 10(11):1777, 2022.

[11] S. L. Wilson and C. Wiysonge. Social media and vaccine hesitancy. *BMJ global health*, 5(10):e004206, 2020.

[12] JustAnotherArchivist. snscrape. Available online at https://github.com/JustAnotherArchivist/snscrape, 2018. Accessed: 2022-10-18.

[13] Radim Řehůřek. Phrase (collocation) detection. Available online at https://radimrehurek.com/gensim/models/phrases.html, 2009. Accessed: 2023-02-20.

[14] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics.

[15] Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246, 2017.

[16] Oscar Araque, Ganggao Zhu, and Carlos A. Iglesias. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346–359, 2019.

[17] Jason S. Kessler. scattertext: a browser-based tool for visualizing how corpora differ. Available online at https://github.com/JasonKessler/scattertext, 2017. Accessed: 2023-03-10.

[18] Ryan J. Gallagher, Morgan R. Frank, Lewis Mitchell, Aaron J. Schwartz, Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds. Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(4), 2021.

[19] Plotly Technologies Inc. Collaborative data science. Available online at https://plot.ly, 2015. Accessed: 2023-04-05.

[20] World Health Organization. A brief history of vaccination. Available online at https://www.who.int/news-room/spotlight/history-of-vaccination/a-brief-history-of-vaccination. Accessed: 2023-05-11.

[21] Yen-Chin Liu, Rei-Lin Kuo, and Shin-Ru Shih. Covid-19: The first documented coronavirus pandemic in history. *Biomedical journal*, 43(4):328–333, 2020.

[22] World Health Organization. Disease outbreak news: Monkeypox – united kingdom of great britain and northern ireland. Available online at https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON381. Accessed: 2023-04-28.

[23] World Health Organization. Update on omicron. Available online at https://www.who.int/news/item/28-11-2021-update-on-omicron. Accessed: 2023-04-29.

[24] World Health Organization. Herd immunity, lockdowns and covid-19. Available online at https://www.who.int/news-room/questions-and-answers/item/herd-immunity-lockdowns-and-covid-19, 2022. Accessed: 2023-05-11.

[25] S. I. Mallah, O. K. Ghorab, S. Al-Salmi, O. S. Abdellatif, T. Tharmaratnam, M. A. Iskandar, J. A. N. Sefen, P. Sidhu, B. Atallah, R. El-Lababidi, and M. Al-Qahtani. Covid-19: breaking down a global health crisis. *Annals of clinical microbiology and antimicrobials*, 20(1):35, 2021.

[26] CNN. Ema declares astrazeneca and j&j vaccine safe despite blood clot concerns. Available online at https://edition.cnn.com/2021/03/18/europe/ema-astrazeneca-vaccine-blood-clots-decision-intl/index.html. Accessed: 2023-04-28.

[27] European Medicines Agency. Ema recommends covid-19 vaccine janssen for authorisation in the eu. Available online at https://www.ema.europa.eu/en/news/ema-recommends-covid-19-vaccine-janssen-authorisation-eu. Accessed: 2023-04-29.

[28] European Commission. Eu digital covid certificate. Available online at https://commission.europa.eu/strategy-and-policy/coronavirus-response/safe-covid-19-vaccines-europeans/eu-digital-covid-certificate_en. Accessed: 2023-04-29.

[29] Lori Robertson. Pfizer documents show vaccine is highly effective, contrary to social media posts. Available online at https://www.factcheck.org/2022/05/scicheck-pfizer-documents-show-vaccine-is-highly-effective/, 2022. Accessed: 2023-05-18.

[30] Centers for Disease Control and Prevention. Key facts about seasonal flu vaccine. Available online at https://www.cdc.gov/flu/prevent/keyfacts.htm. Accessed: 2023-04-28.

[31] BBC News. Covid-19 vaccine: First person receives pfizer jab in uk. Available online at https://www.bbc.com/news/uk-55227325. Accessed: 2023-04-28.

[32] World Health Organization Regional Office for Europe. Sars-cov-2 delta variant now dominant in much of european region; efforts must be reinforced to prevent transmission, warns who regional office for europe and ecdc. Available online at https://www.who.int/europe/news/item/22-07-2021-sars-cov-2-delta-variant-now-dominant-in-much-of-european-region-efforts-must-be-reinforced-to-prevent-transmission-warns-who-regional-office-for-europe-and-ecdc. Accessed: 2023-04-28.

[33] Rachel M. Burckhardt, John J. Dennehy, Leo L. M. Poon, Linda J. Saif, and Lynn W. Enquist. Are covid-19 vaccine boosters needed? the science behind boosters. *Journal of Virology*, 96(3):e01973–21, 2022.

[34] Centers for Disease Control and Prevention. Cdc recommends the first updated covid-19 booster. Available online at https://www.cdc.gov/media/releases/2022/s0901-covid-19-booster.html, 2022. Accessed: 2023-05-12.

[35] US Food and Drug Administration. Fda authorizes booster dose of pfizer-biontech covid-19 vaccine for certain populations. Available online at https://www.fda.gov/news-events/press-announcements/fda-authorizes-booster-dose-pfizer-biontech-covid-19-vaccine-certain-populations, 2021. Accessed: 2023-05-12.

[36] Johns Hopkins Bloomberg School of Public Health. Bivalent covid-19 booster updates. Available online at https://publichealth.jhu.edu/Bivalent-Covid-19-Booster-Updates, 2022. Accessed: 2023-05-12.

[37] Centers for Disease Control and Prevention. Notes from the field: Poliovirus type 2 found in an unvaccinated adult — new york, july–august, august–september, september–october, october–november, november–december, december–january, january–february, february–march, march–april, april–may, may–june, june–july. Available online at https://www.cdc.gov/mmwr/volumes/71/wr/mm7133e2.htm?s_cid=mm7133e2_w. Accessed:2023-04-30.

[38] Nick Andrews, Elise Tessier, Julia Stowe, Charlotte Gower, Freja C. M. Kirsebom, Ruth Simmons, Eileen Gallagher, Simon Thelwall, Natalie Groves, Gavin Dabrera, Richard M. Myers, Colin Campbell, Gayatri Amirthalingam, Mary Edmunds, Maria Zambon, Kevin K. Brown, Susan Hopkins, Meera Chand, Shamez N. Ladhani, et al. Duration of protection against mild and severe disease by covid-19 vaccines. *The New England Journal of Medicine*, 386(4):340–350, 2022.

[39] Centers for Disease Control and Prevention. Administering hpv vaccine. Available online at https://www.cdc.gov/vaccines/vpd/hpv/hcp/administration.html. Accessed: 2023-05-14.

[40] G. Andrade. Medical conspiracy theories: cognitive science and implications for ethics. *Medicine, health care, and philosophy*, 23(3):505–518, 2020.

[41] BBC News. Coronavirus: Bill gates 'microchip' conspiracy theory and other vaccine claims fact-checked. Available online at https://www.bbc.com/news/52847648, 2020. Accessed: 2023-06-12.

[42] P. J. Hotez. America's deadly flirtation with antiscience and the medical freedom movement. *The Journal of clinical investigation*, 131(7):e149072, 2021.

[43] Jill M. Ferdinands, Mark G. Thompson, Lenee Blanton, Sarah Spencer, Lauren Grant, and Alicia M. Fry. Does influenza vaccination attenuate the severity of breakthrough infections? a narrative review and recommendations for further research. *Vaccine*, 39(28):3678–3695, 2021.

[44] US Food and Drug Administration. Emergency use authorization. Available online at https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization. Accessed: 2023-05-14.

[45] JustAnotherArchivist. snscrape: Issue #834. Available online at https://github.com/JustAnotherArchivist/snscrape/issues/834, 2023. Accessed: 2023-06-26.

[46] Sorin Cheval, Cristian Mihai Adamescu, Tilemachos Georgiadis, Mathew Herrnegger, Adrian Piticar, and David R. Legates. Observed and potential impacts of the covid-19 pandemic on the environment. *International Journal of Environmental Research and Public Health*, 17(11):4140, 2020.