Analysing Twitters Sentiment Of Covid-19

B-Thesis

A Pilot Study

Course: MKGB94

Word count: 9032

Written by: Alexander Holm & Alfred Spridzans

2023-03-26

Table of Content

Abstract

The COVID-19 epidemic has had a significant effect on society, and social media networks have served as an important source of information and opinion exchange. In this research, we performed a sentiment analysis of COVID-19-related tweets to obtain insight into public opinion toward the pandemic. We collected 1797 messages using the standard Twitter API and analysed them with R studio and Tableau. The period that we took the data from is 2023-02-20 to 2023-03-02.
We used R studio to perform a more in-depth sentiment analysis with the VADER package, which gave us a more detailed grasp of the sentiment expressed in the tweets. We also used Tableau to examine the frequency of the terms used in the tweets, providing additional insights into the topics that were most commonly discussed. To improve our methods and processes, we relied on earlier related research. We obtained valuable insights into popular sentiment on social media by analysing the text and emoticons used in the messages. Our findings revealed positive, negative, and neutral feelings about the pandemic, with a slight lean toward the negative. The most frequently used words were "amp," "eu," "svpol," "fhm," "vaccinet," and "kina," and emoticons were the needle emoji, the thinking face, and the lack of knowledge emoji, scoring 16, indicating the current debate in Sweden about COVID-19.

Keywords: COVID-19, Twitter, Sweden, Sentiment, Vaccine.

COVID-19-epidemin har haft en betydande inverkan på samhället, och sociala medier har fungerat som en viktig källa till information och åsiktsutbyte. I den här forskningen utförde vi en sentimentsanalys av COVID-19-relaterade tweets för att få en inblick i allmänhetens åsikter om pandemin. Vi samlade in 1797 meddelanden med hjälp av standard-Twitter API och analyserade dem med R studio och Tableau. Den tidsperiod som vi tog data från är 2023-02-20 till 2023-03-02.
Vi använde R studio för att utföra en mer djupgående sentimentsanalys med paketet VADER, vilket gav oss ett mer detaljerat grepp om den känsla som uttrycktes i tweetsen. Vi använde också Tableau för att undersöka frekvensen av de termer som användes i tweetsen, vilket gav ytterligare insikter om vilka ämnen som oftast diskuterades. För att förbättra våra metoder och processer förlitade vi oss på tidigare relaterad forskning. Vi fick värdefulla insikter om populära stämningar på sociala medier genom att analysera den text och de uttryckssymboler som användes i meddelandena. Våra resultat visade på positiva, negativa och neutrala känslor om pandemin, med en svag tendens till det negativa. De mest använda orden var "amp", "eu", "svpol", "fhm", "vaccinet" och "kina", och emoticons var nål-emoji, tänkaransikte och brist på kunskap-emoji, med poäng 16, vilket indikerar den aktuella debatten i Sverige om COVID-19.

Nyckelord: COVID-19, Twitter, Sweden, Sentiment, Vaccin.

1. Introduction and Problem Description

The COVID-19 pandemic significantly impacted society, capturing worldwide attention and concern. In this context, social media platforms such as Twitter have played an essential role in allowing people to communicate their thoughts and feelings about the pandemic and its consequences. This research investigates the connection between sentiment expressed in social media messages about COVID-19 on Twitter in Sweden. This kind of study is essential because of the pandemic's widespread effect on a global scale. Even after the epidemic has passed, the debate over its consequences continues, emphasising the importance of continuing research. By doing sentiment analysis, we can see if there is still much negativity online or if the pandemic still provides fear into people's lives. Users online may have different thoughts about how the state of the pandemic fits our worldview.

There is much to be discussed over the public's perception, so by going into more depth about the public's feelings, we can find expressions of how the word and sentiments affect people's values, ideals and connotations from our daily life through search platforms, like Twitter. As a closure, the covid issue has not been resolved and remains an open question.

Sentiment analysis is critical because it allows academics, organisations, and politicians to understand public opinions, attitudes, and feelings about various subjects, events, or concerns. This understanding is vital for making educated decisions, resolving issues, and recognising the needs and preferences of multiple groups. It remains an open subject in the context of COVID-19 due to the dynamic nature of the continuing pandemic, including the introduction of new strains, vaccination hesitation, and altering public opinions. Sentiment analysis on social media platforms such as Twitter can provide valuable insights into how people perceive the situation, and sentiment analysis can then guide decision-makers in developing effective public health strategies, communication plans, and policy changes to address the challenges posed by the pandemic.

Sentiment analysis has become an increasingly popular study topic in social media. Earlier research has shown that sentiment analysis can provide valuable insights into public opinion and sentiment toward various problems. For example, in the study "Fine-Tuned Sentiment Analysis of COVID-19 Vaccine Related Social Media Data: A Comparative Study" (Melton et al., 2022), sentiment analysis of COVID-19 vaccine-related social media data compared, and researchers discovered that sentiment analysis could provide important insights into public opinion and sentiment toward vaccines.

The authors of the study' Deep learning based sentiment analysis of the public perception of working from home through tweets' (Vohra & Garg, 2023) used deep learning to analyse the sentiment expressed in tweets about working from home. They found that emojis, whereas emojis are pictograms, logograms, ideograms, or smileys, are embedded in the text and used in electronic messages and online sites and contain valuable information about the sentiment expressed in tweets. Furthermore, prior research has demonstrated the significance of emojis in sentiment analysis because they can provide extra information about the sentiment conveyed in social media posts (LeCompte & Chen, 2017; Hauthal, Burghardt & Dunkel, 2019). As the researchers stated, valuable insights into public opinion are not only a means for various administrations to better the well-being of society but also to understand how covid is viewed from an academic standpoint. As a result, this early research could strengthen our investigation case. Furthermore, earlier research has demonstrated the significance of taking into account the number of comments and retweets in social media posts, as they can provide essential information about the popularity and spread of COVID-19-related information (Chandrasekaran, Mehta, Valkunde & Moustakas, 2020; Lamsal, 2021). Moreover, it can give our study meaning from the Twitter posts.

This research aims to thoroughly analyse the sentiment conveyed in social media messages about COVID-19 on Twitter in Sweden and investigate the connection between sentiment and the number of comments and retweets. The study's research topic is: What is the relationship between sentiment expressed in social media posts on Twitter in Sweden regarding COVID-19 and the number of comments and retweets? The goals of this study are first to conduct sentiment analysis on social media posts about COVID-19 on Twitter in Sweden and then investigate the relationship between sentiment and the number of comments and retweets and finally provide insights into public sentiment and opinion about COVID-19 in Sweden.
To better grasp public attitude and opinion toward COVID-19 in Sweden, which can be helpful for policymakers and stakeholders. This report will include the research background and context, theoretical foundation, methodology and motivation, results, discussion, references and appendix.

2. Theoretical Foundation

Sentiment analysis and opinion mining are significant fields of study that have received traction recently, particularly in social media. Sentiment analysis aims to identify and categorise the emotion conveyed in the text, such as social media messages, as positive, negative, or neutral. Emojis, extensively used on social media, have been discovered to provide valuable information about the sentiment conveyed in social media posts and can help learn about other people's opinions on different topics (Yurtoz & Parlak, 2019; Hauthal, Burghardt & Dunkel, 2019).

2.1 In the Age of Big Data, Investigating the Potential of Sentiment Analysis in Understanding Public Opinion and Trends

The theoretical underpinnings of sentiment analysis allow for evaluating context, word relationships, and phrase structures, all required for proper sentiment categorisation.
Sentiment analysis has been a prominent study approach in social media analysis in recent years. The capacity to swiftly and efficiently analyse massive amounts of data has allowed academics to acquire insights into

public opinion and trends (The SAGE Handbook of Social Media Research Methods, 2022). This has been incredibly effective in marketing and politics, where knowing public opinion is critical.

The combination of lexicon-based methodologies and machine learning techniques has yielded promising results in improving sentiment analysis accuracy. Lexicon-based approaches rely on pre-defined sentiment lexicons, which are dictionaries that include words or phrases with assigned sentiment scores. These sentiment scores are then used to estimate a text's overall sentiment. The precision can be more developed by using both lexicon-based approaches with machine learning techniques which R studio provides for our research (Liu, 2012).

As a result of discoveries in deep learning, neural network-based models such as both Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have evolved. These models can capture hidden patterns and representations inside the text, which provides aids in the resolution of some of the challenges that traditional machine-learning methods deal with (Liu, 2012).

Sentiment analysis is a promising and expanding field with the potential for continuous advancement and adaptability to new scenarios. As a result, it has become a necessary tool in today's data-driven environment. It lets academics swiftly and efficiently evaluate enormous amounts of data, revealing insights into public opinion and trends (The SAGE Handbook of Social Media Research Methods, 2022). Overall, sentiment analysis is an exciting method of dealing with data that has the potential to change how we interpret and study human emotions and ideas expressed in text.

Public opinion and sentiment about COVID-19 and its impact are a source of curiosity and concern during the pandemic. Previous research has shown that social media can provide helpful information about public sentiment and opinion about pandemics such as COVID-19 (Mir et al., 2022; Ansari & Khan, 2021). This also implies that social media can help research popular opinion and sentiment during the pandemic. Twitter, for example, provides a wealth of data that can be examined to find patterns and trends in how people discuss COVID-19. Researchers can understand how people in Sweden react to the epidemic, their worries and issues, and how they share their views and feelings by analysing COVID-19-related social media posts.

Furthermore, the statement implies that earlier research has demonstrated the potential of social media data for analysing popular opinion and sentiment regarding COVID-19. This means that the strategy has been tried and tested and that there is actual knowledge and methods to build to explore further the connection between emotions stated in social media messages about COVID-19 on Twitter in Sweden and the number of comments and shares. Overall, this remark emphasises the importance and potential of social media data in answering our research question.

However, sentiment research in social media is full of challenges and limitations. For example, social media users may communicate their emotions in various ways, such as through sarcasm, irony, or negations, making mood analysis more challenging (Ain et al., 2017; Xu et al., 2022). Furthermore, variations in how people communicate emotions in various nations and societies may make cross-cultural sentiment research more difficult (Alhuzali, Zhang & Ananiadou, 2022). This is something that we have to put into consideration when analysing our data and when doing sentiment analysis.

This study will use machine learning methods, including deep learning and sentiment analysis tools like VADER, to conduct sentiment analysis of COVID-19-related social media posts on Twitter in Sweden. The need to provide a complete analysis of the emotion expressed in social media messages and handle the limitations and challenges of sentiment analysis in social media motivated the selection of this method. There are still faults in using these techniques, but we can overcome these issues by performing manual filtering in R so that essential keywords and stopwords do not mix.

This study examines the connection between sentiment expressed in social media messages about COVID-19 on Twitter in Sweden and the number of comments and retweets. The study will examine the sentiment, expression, and emoji usage in COVID-19-related Twitter posts in Sweden.

Emotion is a complicated psychological state associated with various emotions and outcomes. Emotion will be examined in this study using keywords and text analysis. By examining the language used in social media posts related to COVID-19, researchers can gain insights into the emotional responses of individuals in Sweden towards the pandemic.

The sentiment is the positive, negative, or neutral expression of feeling in writing, such as social media posts. Natural language processing and machine learning techniques are used in sentiment analysis to determine sentiment. Researchers can learn how people in Sweden react to the COVID-19 pandemic by studying sentiment in COVID-19-related Twitter tweets.

Lastly, emojis in social media communications communicate emotions and opinions. Emojis are pictorial depictions of feelings widespread in social media communications. Researchers can gain insights into how people in Sweden communicate their feelings and opinions about the epidemic more visibly and naturally by studying emoji use in COVID-19-related Twitter messages.

The connection between social media opinion and real-world events, such as the COVID-19 epidemic, is complicated and can be viewed from different perspectives. While social media can bring valuable insights into public opinion and emotion, it is essential to remember that social media messages are affected by variables other than the actual event.

The language used in media reports on the pandemic and focusing on specific parts of the crisis can give an understanding of how people view and react to the situation. Media coverage of a specific incident, such as the COVID-19 pandemic, can, for example, influence popular opinion and sentiment on social media. Likewise, political discourse regarding the pandemic might influence public opinion and sentiment on social media. Political leaders and organisations responding to the crisis, including their rhetoric and policy choices, can influence how people feel and talk about the epidemic and how they express themselves on social media (Vaccari et al., 2015).

Personal events and individual differences can also impact COVID-19-related social media messages. Users who have experienced the pandemic's negative effects, such as sickness, job loss or even the loss of a close one, might communicate different connotations and emotions on social media than those whom a significant effect has not directly impacted.

These variables that have a negative impact on the public add to the complexities of the association between social media sentiment and real-world events such as pandemics. While social media can provide valuable insights into public opinion and sentiment, it is critical to evaluate these results critically and consider the various variables that may be affecting COVID-19-related social media posts.

The research project aims to comprehensively analyse the sentiment conveyed in COVID-19-related social media posts on Twitter in Sweden. This research will be carried out with the help of a combination of machine learning and sentiment analysis tools. The need to comprehensively analyse the feelings communicated in social media communications and the difficulties and constraints connected with sentiment analysis in social media motivates these techniques.

In times of disaster, such as the COVID-19 epidemic, social media has surfaced as a powerful tool for sharing views and feelings. The wealth of user-generated social media content data can be used to evaluate popular sentiment and opinion. However, sentiment analysis in social media is complicated due to the casual character of social media content, which frequently includes sarcasm, irony, and slang. Machine learning methods and sentiment analysis tools can help resolve these issues by automating the processing and analysis of social media data.

The primary objective of this research is to learn more about popular sentiment and opinion in Sweden regarding COVID-19. The findings from this study could help policymakers, public health officials, and researchers better comprehend the people's views, opinions and ideas about the pandemic. This understanding can be applied to public health initiatives, communication strategies, and policy decisions. Furthermore, the study may serve as a foundation for future research on the changing general perception of COVID-19 in Sweden. Overall, this study can significantly add to public discourse and decision-making in Sweden regarding the COVID-19 pandemic.

2.2 Frequency

We determine the word count (frequency) of all the words to comprehend what has been said about COVID-19 on Twitter. This will help us understand what is important and discussed in this case.

"Word frequency norms Estimates of how often words are encountered based on counting their occurrences in representative corpora." (De Groot et al., 2017, p. 243).

Therefore word frequency (word count) analysis is a helpful instrument for finding patterns and trends in the language used on a specific website. The frequency of words and emojis in a text can provide insight into word usage trends and the general mood communicated (Rajput. et al., 2020).

3. Methodology and Motivation

To conduct this study, we used a data-collecting strategy that involves gathering information regarding the COVID-19 epidemic in Sweden through the social media network Twitter. We collected the data via scraping tweets, a widely used strategy for obtaining and getting information from the site (Janetzko, 2017). Our teacher gave the Twitter data after obtaining many tweets through a default API containing text, phrases, emoticons, and hashtags linked to COVID-19 in Sweden.

The capacity of Twitter to deliver real-time updates and insights into the current crisis inspired the adoption of the network as a source of information regarding the COVID-19 outbreak in Sweden. Adopting Twitter scraping as a data-gathering strategy allowed us to access a significant number of messages relating to the issue at hand, giving us a complete and diversified dataset.

The API usage was to collect information on COVID-19 and related keywords focusing on the geography of Sweden. The API returned a list of tweets with text, keywords, emoticons, and hashtags relating to the issue, allowing us to acquire a significant amount of relevant data fast and efficiently.

Finally, the data collecting approach we used for this study was developed to acquire information regarding the COVID-19 epidemic in Sweden from Twitter utilising scraping tweets and a default API. This method gave us access to a diversified dataset, which offered us insights into the current situation in Sweden.

3.1 Choice of Method and Motivation

Many significant considerations influenced the decision to use Twitter scraping as a data-collecting strategy. For starters, the platform's ability to generate a large number of data fast was a crucial factor in our decision-making process. This was especially critical given the COVID-19 pandemic's fast evolution and the necessity for real-time information on the situation in Sweden.

The platform's default API was another aspect that impacted our decision to adopt scraping Twitter as a data-collecting strategy. The API allows for a simple and quick method of obtaining data, which was critical for our research, given the time-sensitive nature of the data we were gathering. Furthermore, as Janetzko (2017) explained, the API ensures that the data provided is current and relevant to the study subject. Furthermore, using Twitter's API protects the privacy and confidentiality of the data and participants. According to Townsend and Wallace (2016), the dataset obtained through the API is anonymous and does not include any personal information about their comments on Twitter. This is essential in ensuring that the privacy rights of those participating in the research are respected and maintained.

Finally, the decision to use Twitter scraping as a data collection method was influenced by several key factors, including the platform's ability to generate a large volume of data quickly, the straightforward and fast approach to data retrieval provided by the platform's default API, and the protection of the data and participants' privacy and confidentiality.

3.2 Research Design

The authors made a concerted effort to select the sample size in this study to guarantee that the results were representative of the population and appropriate for the sentiment analysis methodologies utilised. The study included all Swedish Twitter users who tweeted about the COVID-19 epidemic from 2023-02-20 to 2023-03-02.

To confirm the sample's representativeness, the researchers used R Studio's scraping function to collect 1797 posts from this group. This sample size was chosen for the study because it was large enough to offer a complete picture of the attitudes expressed by Swedish Twitter users concerning the COVID-19 outbreak while still being manageable for the sentiment analysis techniques utilised.

The sample was collected in a systematic and thorough manner to reduce the likelihood of bias and to guarantee that the results obtained were representative of the more significant population. The scraping tool in R Studio enabled the efficient and effective capture of data, which was subsequently used for sentiment analysis in the study.

Finally, determining the sample size was a critical component of the study since it ensured that the results were representative of the population and appropriate for the sentiment analysis methodologies utilised. The researchers used R Studio's scraping tool to collect 1797 tweets from Swedish Twitter users who had tweeted about COVID-19 in order to conduct sentiment analysis in the study.

3.3 Data Pre-processing and Cleaning

Prior to doing the sentiment analysis, it was critical to guarantee that the data acquired was of high quality and trustworthiness. The data went through a pre-processing and purification stage by cleaning the word gathered from Twitter. The goal of this step was to delete any unnecessary content, such as Links and mentions, and to repair any data errors.

Normalising the data was a crucial step in the pre-processing and purification stage to guarantee that the text was uniform and easy to analyse (Garg & Sharma, 2022). This entailed ensuring that the text was formatted consistently and that any differences in the representation of the same terms were standardised.

The investigators utilised a Rstudio function to delete phrases that were judged unrelated to the analysis during the cleaning phase. This was accomplished by employing the Tidyverse software (Tidyverse, n.d.), which supplied a pre-processed list of generic phrases deleted from the data.

Finally, the pre-processing and purification stage was critical in the analytical process since it ensured that the data acquired was of high quality and reliability. Additionally, the data were standardised to ensure that the language was consistent and easy to analyse. The cleaning procedure included making a list of words that did not match the analysis and using an R studio command to remove the generic phrases that had been pre-processed by the tidyverse package. After the dataset was conducted in R studio, the next phase was to clean it in Tableau for the frequency word cloud and bar chart. The data was cleaned manually in Tableau to eliminate the common words so that they did not affect any graphs.

3.4 Sentiment Analysis Techniques and Tools

VADER is a sentiment analysis software designed exclusively for social media data (Balli et al., 2022). This study's sentiment analysis was carried out using Rstudio and the VADER program. The method is based on a lexicon that was created to assess the emotional values of a specific text.
The VADER program used a vocabulary of terms and associated sentiment ratings to estimate the sentiment of COVID-19-related Twitter posts in Sweden. The method took into account the tweets' positive, negative, and neutral attitudes, allowing for a complete evaluation of the views expressed by Swedish Twitter users concerning the epidemic instead of the Turkish Twitter users that Balli (2022) was researching.

To gather all of the sentiment evaluations in the study, the VADER program was employed in combination with the R studio platform, and the VADER package (Hutto & Gilbert, 2014) was used. The program enabled rapid and effective sentiment rating determination, offering a complete overview of sentiments voiced by Swedish Twitter users concerning COVID-19.

Finally, the sentiment analysis in this study was carried out with the VADER program, which is a sentiment analysis tool designed exclusively for social media data. The method was used to assess the sentiment of COVID-19-related Twitter tweets in Sweden, taking into account the positive, negative, and neutral attitudes stated in the tweets. The sentiment ratings were determined efficiently and effectively using the VADER package in R studio, offering essential insights into the attitudes expressed by Swedish Twitter users concerning the epidemic.

3.5 Tableau

To analyse and display the data in a straightforward and concise way, the researchers used a range of coding tools. Tableau was one of the tools used to improve the visual depiction of the findings. The researchers were able to create visually stunning and informative diagrams and visuals using Tableau, which served to convey the key results of the study in a more approachable manner. The data for the analysis was obtained from R Studio and carefully cleaned and prepped for analysis by the research team. The study also aimed to identify Sweden's most frequently used words and emojis linked with COVID-19. The study produced graphs and graphics showing the most commonly used words and emojis.

The research used Tableau to show the sentiment analysis data in a visually appealing manner. A horizontal bar chart was constructed for the text analysis by ranking the data from highest to lowest. The graphic showed the number of times specific terms appeared in columns and the words themselves in rows.

In Tableau, we chose to go with the divergent green-blue palette to give the amount of occurrences a colour representation to improve the visual clarity of the chart.

In addition, the researchers constructed a word cloud to visualise the most frequently occurring terms in the dataset. They created a new document, converted the marks to text, and entered the number of occurrences twice, once in the size bar area and once in the colour part. The researchers also utilised filters such as MAX(N) and Word to exclude the most frequently occurring terms.

The researchers utilised Tableau to display the results of the emoji analysis in addition to the text analysis. Before importing the data into Tableau to build an emoji bar chart, it was first sorted in R studio, cleaned and adjusted for Tableau. The chart was designed as a horizontal bar graph, with the emoji table in the rows and the number of occurrences in the columns. To indicate the values on the bars, the investigators appended the number of occurrences to the markers.

3.6 Emoji Analysis and Their Relationship to Sentiment

In addition to sentiment analysis, the researchers looked at the use of emojis in postings and their link to sentiment. Using Rstudio, the investigators discovered and classified the emojis used in the submissions. The submissions' sentiment scores were then compared to the use of emojis to see whether there were any trends or relationships between the two. This study shed light on the association between the sentiment represented in the posts and the use of emojis.

A significant component of the study was the investigation of emoji usage and its link to sentiment since it gave a greater knowledge of how emojis are used to communicate emotions and feelings on social media. The findings of this study could have consequences for how other research sentiment analysis is produced on social media data, as well as the significance of capturing emojis in sentiment analysis. Implementing the emoji sentiment gives the topic a grasp of what is communicated in today's Tweeting world.

Finally, an essential aspect of the study was to appraise emoji usage and its correlation to sentiment. The research recognised and classified the emojis used in the postings and compared the posting sentiment ratings to the use of emojis to see whether there were any trends or connections between the two.
The findings of this study gave valuable insights into the link between sentiment and emoji usage, emphasising the relevance of including emoticons in sentiment research.

3.7 Analysis of Comments and Retweets

The researchers also looked at the link between the mood expressed in the posts and engagement measures like the amount of comments and retweets for each post. Tableau was used to create visualisations of the distribution of comments and retweets for posts with varied sentiment scores.

The investigation of the link between sentiment and engagement measures yielded helpful insights into the elements that drive social media participation. Our investigation also emphasised the importance of mood on interaction, with posts expressing positive or negative attitudes obtaining more comments and retweets than ones with neutral sentiments.

This analysis' findings have ramifications for social media marketers and researchers because they give a better understanding of the link between sentiment and engagement measures. The researchers acquired a more profound knowledge of the elements that drive engagement on social media by evaluating the link between sentiment and engagement indicators, giving significant insights for future research.

Finally, an essential part of the study assessed the link between sentiment and engagement measures. The researchers used Tableau to construct visualisations of the distribution of comments and retweets for posts with varied sentiment scores, offering significant insights into the link between sentiment and engagement measures. The findings of this study have significant significance for social media marketers and researchers, emphasising the effect of sentiment on engagement measures.

3.8 Ethical Considerations

During the study, the investigators kept ethical considerations and best practices for social media research and data protection in mind. This included ensuring that the data acquired was anonymous and contained no personally identifying information (Townsend & Wallace, 2016).

The investigators prioritised the security of the participants and the data's privacy and confidentiality. Using anonymous data and the lack of personally identifying information ensured that participants and their information were safe from damage or misuse (Townsend & Wallace, 2016).

The researchers followed best practices for social media research and data security, ensuring the study was ethical and responsible. This kind of method not only protects the participants and their data and increases the results' trust and validity (Townsend & Wallace, 2016). Finally, the investigators in this research adhered to best practices for social media research and data security, ensuring that the study was carried out ethically and responsibly, increasing the reliability and validity of the results. However, one can also argue that, by accepting the terms of Twitter and cookies, they already accepted the terms of service by using the platform, meaning their data is legally collected. It is complicated to verify that collecting data by a user who accepted the terms agreed to be used in research. The reasoning is that most people do not always know that by accepting the cookies, the companies that withdraw information about the user have all the legal right to do so. It is an open discussion in research ethics because most people do not know what they agree to when they accept cookies. This ethical consideration does not only applies to Twitter's platform but also most of the common platforms online. For example, Snapchat is used in Sweden, without accepting the terms of service, the user is prohibited from creating an account. This makes it complicated for the users because without consenting to the terms of service, they can not be connected with society.

4. Results

4.1 Frequency - Text

The research examined tweets related to the COVID-19 pandemic in Sweden to grasp public opinion toward the pandemic better. Tableau and R studio evaluated tweets data samples from February 20, 2023, to March 2, 2023, to determine the most frequently used words and sentiments linked to COVID-19 in Sweden.
The analysis results of Tableau showed that the most frequently used words in tweets related to COVID-19 were in Sweden, which is relevant because of our research question. The top of the list was "amp", "eu", "svpol", "fhm", "vaccinet", and "kina".
Overall, these results provide significant insights into the subjects and concerns that Swedish Twitter users are most concerned about when discussing COVID-19. The research offers a foundation for further investigation and analysis of public sentiment toward the pandemic in Sweden by finding the most frequently used words and topics.

The Tableau word cloud in Figure 1 depicts the most commonly used terms in tweets about COVID-19 in Sweden. The cloud gives an overview of the most commonly used words, stretching from a word count of 28 up to the highest of 169. The word cloud show the most common word by colour and size; the more blue the colour is, the higher frequency that word has. The darker green the words become, the less frequent the words get.

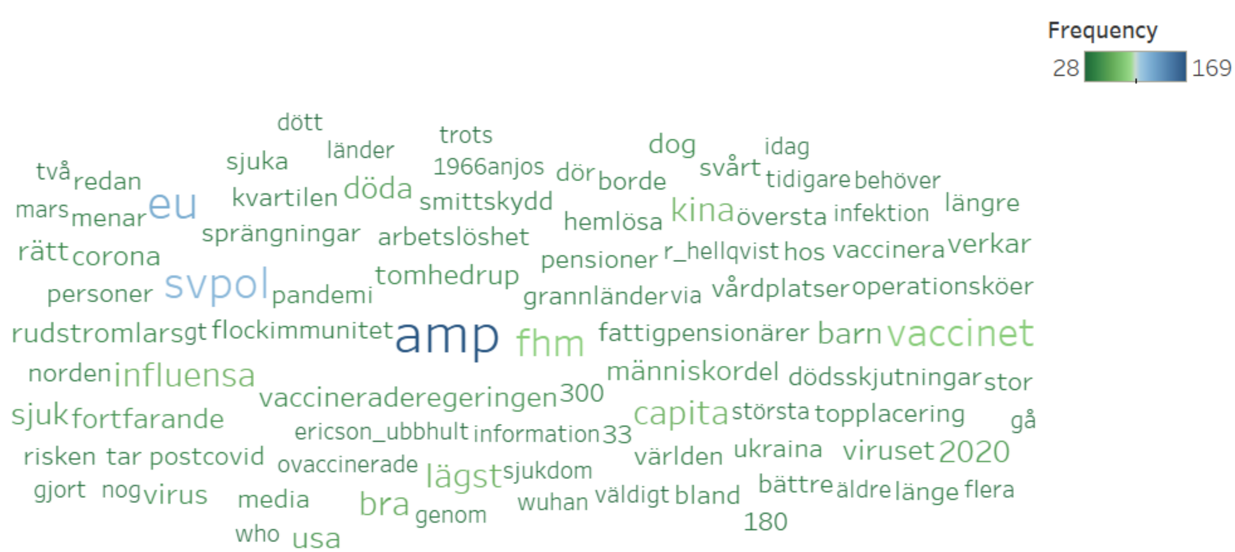


Figure 1 - Wordcloud shows the Frequency of Word Usage

Figure 2 shows a horizontal bar chart produced by Tableau that also shows the frequency of each term used in tweets regarding COVID-19 in Sweden. This bar chart provides a more visual representation of the most frequently used terms.

According to the research, Sweden's most commonly used terms in tweets about COVID-19 were amp, eu, svpol, fhm, vaccinet, and kina. Tableau's word cloud and horizontal bar chart depicted the most frequently used terms, making comprehending trends and patterns in tweets about COVID-19 in Sweden more straightforward.

The analysis findings can be used to gain insight into the pandemic's debate and views on Twitter in Sweden. The most commonly used word was "amp," scoring 169, which could be from the Bergman - AMP research organisation in Karolinska Institutet (*Research group bergman - The AMP-group, 2023*), which focuses on the topic covid-19. "Eu" was also at the top of the list, with a score of 120, indicating European Union. Another frequently used term, "svpol", with a score of 114, seems to be a political hashtag used on related political discussions on Twitter. A study done by Lorentzen found the hashtag #svpol in 2012, which functions as a Twitter platform for debating Swedish politics and society. According to conversations analysed, this hashtag is mainly used for political criticism between users with opposing views who seek to question each other's positions in various ways. Respondents in the research voiced dedication to democratic ideals, a love of public debate, and a common interest in social and political problems (Lenstierna, 2019). FHM(https://fhm.se/) had a high frequency in the discussions on Twitter, which is a website for Swedish people to give advice on how to tackle the different challenges in society in relation to covid-19. The term "vaccinet," a Swedish word for the vaccine, has a frequency rate of 84, indicating that there is a continuing discussion about this word. The frequency of these words is still focusing on the pandemic, which means the platform Twitter comments within the field of covid still raising questions on how covid affects the society today.

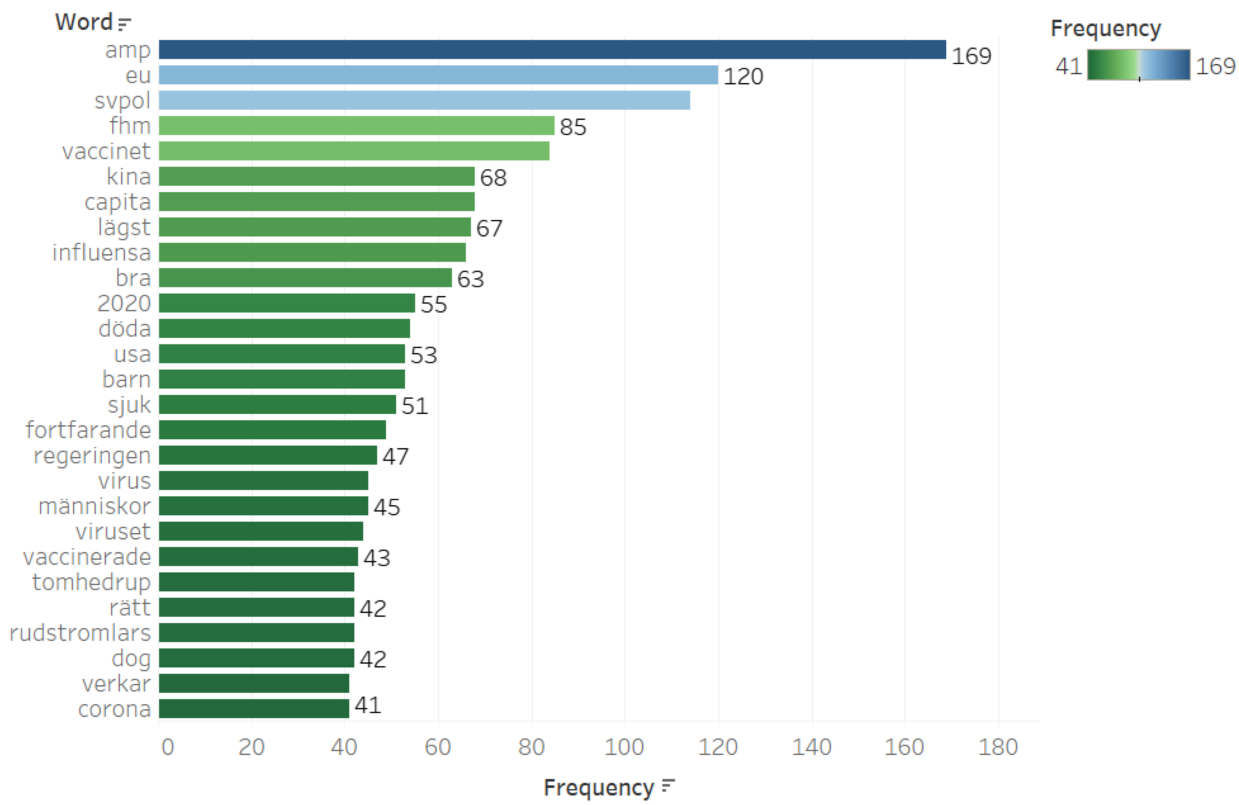


Figure 2 - Bar chart shows the Frequency of Word Usage

In figure 2, we used the same model as figure 1, where the colour goes from dark blue to green dark green. By looking into the visualisation, we can see that the graph's top words are blue, and the more frequently the word is used, the darker(blue) it becomes. When we see the word corona, which is a the bottom, it shows the darkest green colour for the chart. This means that the word corona has a minor frequency in the graph; the lighter the green colour becomes, the more frequent the word is.

4.2 Sentiment analysis - text

The offered example displays a summary of sentiment scores and the number of occurrences for each score. It resembles a tabular representation of the Vader algorithm's sentiment analysis results.

The sentiment scores range from -1 to 0.6, with lower values indicating negative sentiment, higher values suggesting positive emotion, and lower values indicating neutral sentiment. The count column displays the number of times each sentiment score appears in the text.

According to the sample, the bulk of the sentiment ratings (60 out of 80) appear to be neutral (0), with 11 occurrences of negative sentiment scores (-1) and only a few events of positive sentiment scores (0.4 and 0.6). It is vital to remember that elements like the language employed, the text's context and the analysis technique's limits can all impact the outcomes of sentiment analysis.

Twitters found each post's mood using a sentiment analysis tool. Figure 3 shows a bar plot that outlines the findings of a sentiment analysis performed on social media messages about COVID-19 in

Sweden by using R. It was evaluated on a range of -1 to 1, with -1 representing a highly negative sentiment, 0 representing neutral sentiment, and 1 representing a highly positive sentiment.
Figure 3 indicates that most posts (60 out of 78) were neutral, 11 were extraordinarily negative, and only four were mildly negative. Positive sentiments ranged from 0.05 to 0.6 on the sentiment measure in a few posts. Figure 3 offers valuable insights into the general attitude conveyed in COVID-19-related social media messages in Sweden. Furthermore, It also provides a foundation for future research into the relationship between stated emotions and the frequency of comments and retweets on COVID-19-related posts in Sweden.

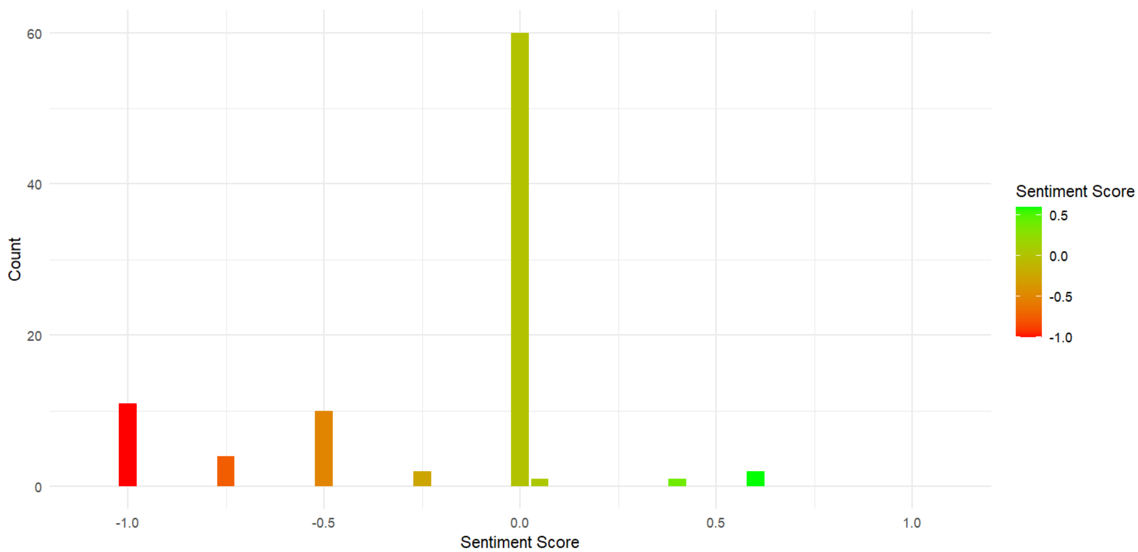


Figure 3 - The bar chart shows the sentiment score of the Twitter text
This bar chart shows the sentiment score of all the text. The visualisation goes from red, yellow and green, where the redder the chart becomes, the higher the negative sentiment score appears. Yellow in this graph is the colour of neutral, the more centred the bar is on the chart, the more neutral the score becomes. Lastly, the green colour shows a positive sentiment score, and it goes from the neutral colour yellow to a more green colour depending on how much the sentiment score gets closer to 1.

4.3 Frequency - Emoji 😊

The Tableau program was used to investigate the use of emojis in textual data. The goal was to determine the most commonly used emojis and their associated ratings. The most frequently used were the needle emoji (💉), the thinking face (🤔), and the lack of knowledge emoji (🤷), scoring 16. The laughing emoji (😂) was discovered to be the second most commonly used, with a score of 14, implying that people use comedy as a coping strategy or to communicate positive feelings. This indicates that people are showing uncertainty or confusion in their textual interactions.

With a score of 11, the male sign emoji (♂) was also a frequently used symbol, indicating that gender may be a subject of conversation or thought in the examined messages. Overall, the research casts light on the prevalence of particular emojis in textual interactions and their possible effects on individual emotions and views. This emoji topic is discussed more in the discussion part of this research.

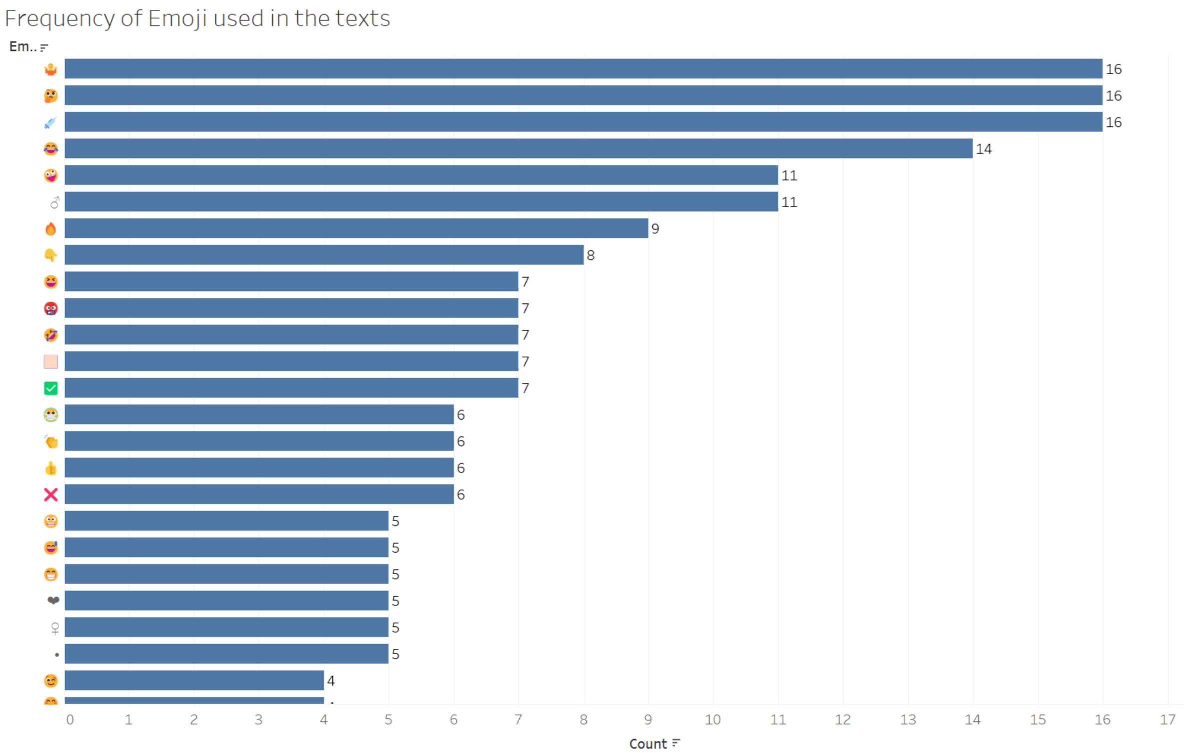


Figure 4 - The bar chart shows the Frequency of Emoji Usage

In figure 4, the chart of the different emojis descending from the highest most used emoji to the least used emoji at the bottom of the graph.

4.4 Sentiment analysis - Emoji 😊

Figure 5 shows a bar plot that gives significant insights into the sentiment analysis of COVID-19-related social media posts in Sweden. The plot displays the sentiment score for each of the emotion groups examined in this study, providing a clear comparison of the various sentiments conveyed in social media posts. According to the bar plot, the sentiment category with the highest sentiment score is neutral, suggesting that the majority of the social media posts examined in this study conveyed a neutral sentiment. This indicates that people in Sweden are adopting a more balanced and objective approach to addressing the pandemic on social media rather than showing strong positive or negative feelings in their COVID-19-related social media posts.

What is the relationship between sentiment expressed in
social media posts on Twitter in Sweden regarding COVID-19 and the number of comments and retweets?

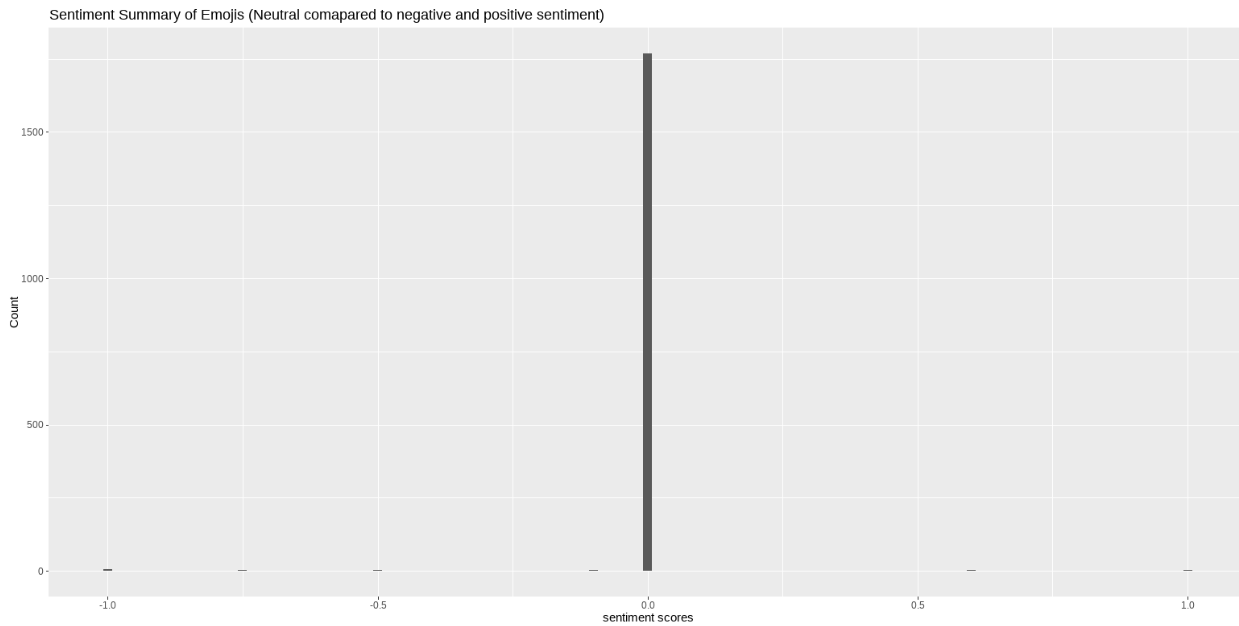


Figure 5 - In this bar plot, we can see that the neutral sentiment is the highest.
The bar plot accurately depicted the overall emotion score, which was neutral. This visualisation gave valuable insights into the opinions stated by Swedish Twitter users concerning COVID-19, underlining the fact that the bulk of the sentiments conveyed were neutral. Also, in the chart, by looking deep into the small numbers near the negative sentiment that appears to the left of the graph and the positive score that appears right to the graph, that shows inside the graph how many negative and positive sentiment score emojis had compared to the neutral sentiment score.

Figures 5 and 6 show the results of a study into the relationship between sentiment and emojis in COVID-19-related social media comments in Sweden. The figures show a summary of the number of posts for each sentiment score as well as the number of emoticons used in each post, which provides useful insights into how people communicate their feelings and use emojis to enhance sentiment in their social media communications.
Figure 5 indicates that most of the social media posts examined in this study (3538 out of a total of 3564) expressed neutral emotions, with the remaining posts showing a mix of positive and negative emotions. Meanwhile, Figure 6 shows that negative sentiment ratings and emoticons were used more frequently in COVID-19-related Twitter posts in Sweden.
These results cast insight into the relationship between sentiment and emojis in COVID-19-related social media messages in Sweden. According to the findings, emoticons can be used to enhance sentiment analysis results and communicate feelings in social media communications. Researchers can obtain a more nuanced grasp of the emotional tone of social media communications linked to COVID-19 in Sweden and how these feelings are conveyed through emojis by combining sentiment analysis and emoji analysis.

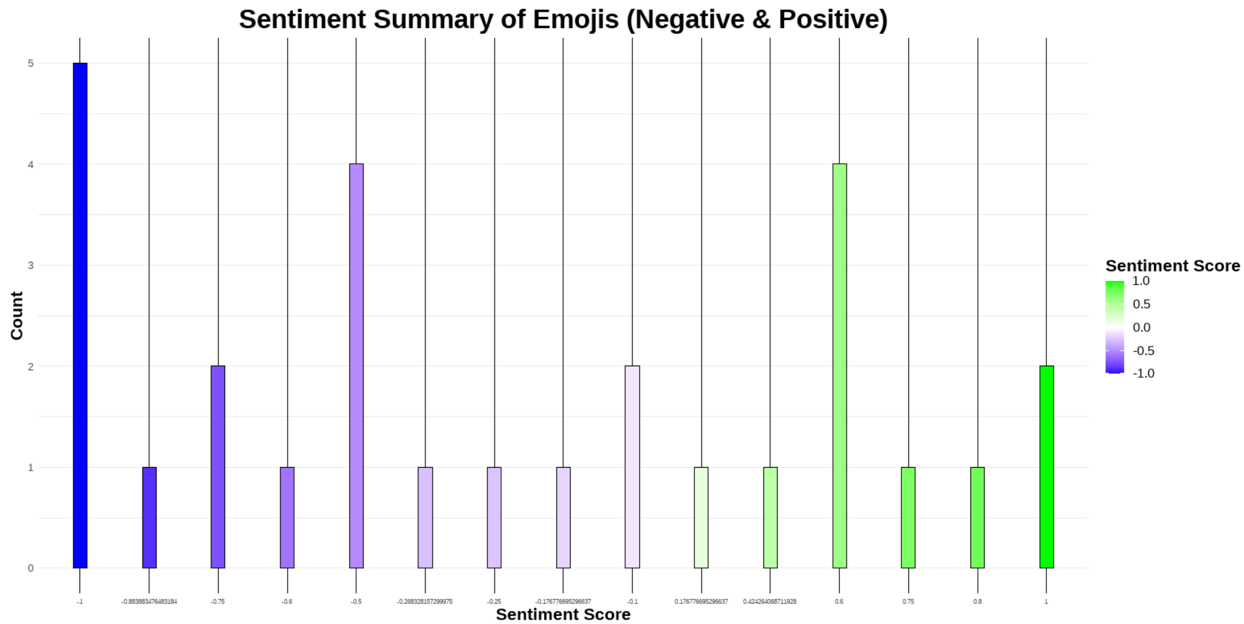


Figure 6 - bar chart that shows the positive and negative sentiment of the emojis on Twitter.

The sentiment analysis on emojis results were visually depicted as a bar chart. The positive and negative sentiment summary of the emojis used in the postings is shown in this bar chart. The graphic clearly showed that the negative sentiment outweighed the excellent mood. The visualisation of figure 6 shows the negative sentiment score going from blue(negative sentiment score) to green(positive sentiment score). The more the sentiment score becomes positive, the green the colour gets; it has the same colour visualisation as figure 1 and figure 2.

4.5 Summary

The results could clearly show a correlation between the sentiment of the text towards covid-19 in Sweden, with 1797 comments and retweets; most of the sentiment was neutral and was still an ongoing topic on the Twitter platform. By looking into the tweets, we could analyse both the text and the emoticons on the platform. The sentiment results of both emojis and text were almost identical, the values of the score were most neutral, but there was more negative sentiment than positive sentiment on the pandemic. However, the results also show some positive sentiment towards public opinion in Sweden. With the help of the frequency, we could also see that the words who was most frequently used were a question about the topic concerning the pandemic. They bring up AMP (*Research group bergman, 2023)* most in the frequency, with a score of 169. This group conduct research towards covid-19. This showed us that people still have an interest in discussing the subject covid-19 online and are interested in how far the research has come in understanding the pandemic. During a global health crisis in Sweden, this study can give useful insights into public opinion and the effectiveness of communication tactics.

This answers our research question: What is the relationship between sentiment expressed in social media posts on Twitter in Sweden regarding COVID-19 and the number of comments and retweets?

5. Discussion

From today in Mars 2023, there is hardly any discussion about COVID-19 in the news, which is intriguing because it is still a pandemic but not as deadly as it was a few months ago. As a result, this research is fascinating to learn about what is said and discussed on COVID-19 today. Still, this is a more directed sentiment analysis towards the population of Sweden.

The emoticons of the results included all kinds of expressions in the text. As mentioned previously, there is a difference in the frequency of emojis and their usage in the text. The results could either depend on the availability and accessibility of their choice of words. Implementing emojis in the text could also give a higher meaning and strengthen the user's message. It is important to remember that emoticon's meaning differs from person to person. Depending on their personal experiences, societal background, and circumstances, different individuals may perceive the same emoticon differently. As a result, it is critical to use emoticons correctly and with caution, particularly in professional or formal situations where they may be inappropriate.

From our interpretation, the top of the text's frequency indicates neutral sentiment. Where amp, eu, svpol, fhm, vaccinet, and kina sits on the trophy. These appear to be locations, organisations, or common words. These are the most common and can convey an indifferent or favourable tone based on the context of the text. As an illustration, consider the nation of Kina. If the text blames Kina for beginning the pandemic, this correlates to a negative text, implying that the term kina is a negative sentiment in this reading. As a result, these top words could have a context to either have a more positive or negative sentiment.

5.1 Previous research

As stated in the introduction, the prior research that we used for our study helped us by providing a clear overview of how we could tackle our research topic. That said, for example, the material provided by Melton et al.(2022) is critical to our research question because it demonstrates the ability of sentiment analysis to provide insights into popular opinion and attitudes toward COVID-19 in social media posts on Twitter and Reddit. According to Melton et al.(2022), the sentiment on 2020-2022 on Twitter was the most negative, while Reddit had a more positive general sentiment. Therefore can help if this is needed to compare results or methods with the different years.

Different tools were appropriate for this comparative study, and Vohra et al. (2023) used the sentiment tool VADER in their study of a sample of 450,000 tweets. This is something that we also included in our research.

5.2 Future studies

Additionally, as written before, this research can be beneficial to find hidden patterns in social media by not only focusing on the continent Sweden and magnifying the view by including different countries that can widen the research area. By looking into another culture, the sentiment score might change. Sweden was a country with very low restrictions both when the pandemic was at its highest peak and now, in the year 2023, the pandemic almost has no restrictions at all. This can be compared to countries that had a very strict policy for vaccinations and laws in other countries to see what the sentiment on covid-19 has in comparison.

Furthermore, looking into other platforms can contribute to a larger dataset and also see if there is any comparison between platforms' sentiments. By doing studies like this, corporations can gain insight into what the public online feels about certain topics. The time frame can also be interesting to research if the messages were written at a certain time frame of the day does change the sentiment of the audience of the platforms.

Finally, one thing this analysis could be made to gain even more information about the sentiment would be to analyse the favourable comments and retweets to gain a higher understanding of what kind of topics are the most interesting and what sentiment they provide.

5.3 Limitations

When we ran the VADER sentiment analysis, we ran into issues with the content of the Twitter posts. Some of the words written in the dataset had various misspellings, and they were the most frequently used words, which became a huge issue for us when we used the cleaning function in R studio because we had to clean out some of the misspellings manually. According to Hutto and Gilbert (2014), the system VADER is not flawless; it still has linguistic constraints, and the irony is a factor that can disrupt the analysis.

We cleaned it because many of the terms were frequently misspelt. Regarding humanity's slang and a dearth of proper spellings, R studio's various products have constraints. Even though the VADER package's algorithms have a method for understanding some of the informal words, it does not encompass everything. That is why we needed to start thinking about how much work the research should put into this portion of the undertaking.

The packages needed updates, which was not only time-consuming but also a way of approaching the coding from different views. That is why we chose to work in combination with R studio and Tableau. In Tableau, we could use the CSV files that were already filtered and then easily conduct charts with a higher rate of work and less time trying to build nice graphs in R studio.

When it came to data scraping, the researchers first had difficulty obtaining a default API key; without an API key, there is no link to any data, which is why we required assistance from the supervisor to collect information from Twitter. R studio's error codes are frequent, and it is critical to perform the coding perfectly. Everything breaks apart when there is a missing comma or an incorrect decimal, and the programmer must find a method to interfere with the problem.

The visualisations of R studio had some limitations for the project; The graphs need the correct amount of time to be able to have everything into an aspect of what is the most needed in the designs of the visualisations. If something is missing or one file is wrong, the dataset will then be conducted falsely, which would make the analysis unreliable.

As mentioned before, it is important to research the terms with more context to know if it is the correct meaning. Moreover, this is important to know to set the correct sentiment for each term. A word on Twitter can give different meanings, which we had to consider when analysing the top words. For example, the word "fhm" can stand for different things, but in this context, we could dig deeper into the hashtag by searching on Twitter and the R file to know what has been talked about around this term.

5.4 Conclusion

To conduct sentiment analysis, the VADER program for R studios Library was a good way of dealing with a large dataset. It helped our research to conduct the project easily because it already has a built-in lexicon (Hutto & Gilbert, 2014). With this method, the analysis did not need to watch over every word in the Twitter comment section; instead, using a CSV file that had all the data and then used the library Sentimentr to code the text to gain a sentiment on almost every word (Quinn et al., 2020). That is why this project chooses this method for quantitative research.

The validity of the research could have been the approach from a large mass of data; this would have meant that we needed higher access from Twitter's API. The default access is between ten days of content from Twitter, but with access to the developer platform(https://developer.twitter.com/en) on Twitter's website, the collection of data would have been drastically increased. This would have validated our approach even more and given a large time frame of data for our project, which meant more to analyse.

Even if we got a sentiment score from R studio, we could have cleaned it manually even further, leading to a bigger word cloud and more precise graphs.

Seeing such a high neutral sentiment in the analysis was interesting because it shows that people do not feel bad about our situation after the pandemic has become less relevant.

When we used R studio, the filtering became hard because it should have removed most of the Swedish stopwords, but after opening the file into Tableau, the words that were supposed to be filtered out were still there. Combining both R studio and Tableau made the cleaning much more simply because you can remove irrelevant words within the graphs in Tableau's program. That is why we utilised both programs to have a broader view of the data.

The analysis showed that the emoticons and text shared the same sentiment. By conducting the results, the analysis could confirm that the data of Twitter covering the sentiment on covid-19 was overall neutral and draw conclusions to the report that Twitter does not cover the same amount of negative sentiment score as previous research. The data from Melton et al.(2015) indicated that there was much more negative sentiment on the matter from the years 2020 - 2022 compared to this report, which almost got every sentiment score neutral. The reason why the result differs might be because the pandemic is no longer the same threat to society, instead being treated as a standardisation the same as the flu. This can be defined as a more normal disease and does not has the same impact on society anymore. With more time, technology advances, and the public grasp the difference between dangerous and common diseases.

5.5 Conclusion R Studio Design

When we conducted the coding in R studio, we got different visualisations. Finding an excellent suite of our results of the sentiment scores was challenging. A trail of different views needed to be conducted to get a good view of the different charts for the study.

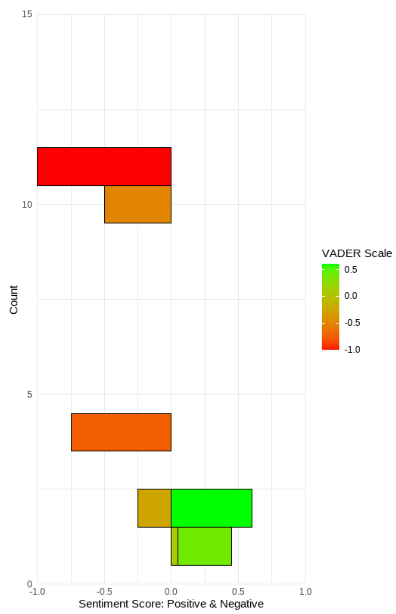
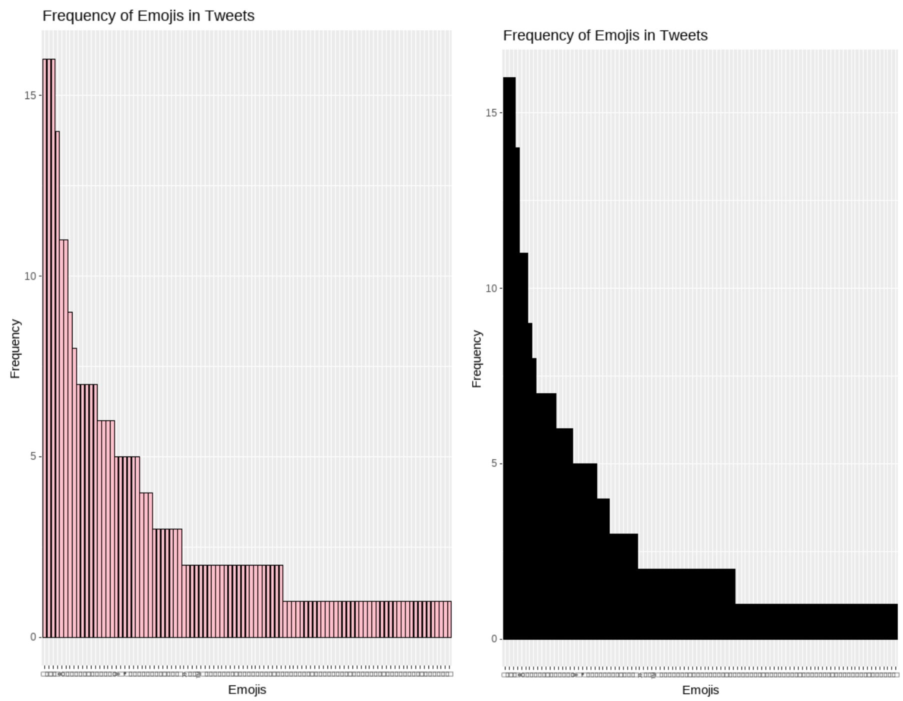


Figure 7

In this figure 7, which was our first test chart it was difficult to understand our data, which is why our study needed the right design for the bar chart.

That is why we choose to keep figure 6 instead. It feels easier to read for the eyes and gives the correct understanding of our results.



Figures 8 & 9

This chart had too many errors, The chart did not provide any insight into what emojis' frequency was. The colour was not correct, and the texture package could not find how the emojis' looked. Because of this, we used the Tableaus graph instead(figure 4).

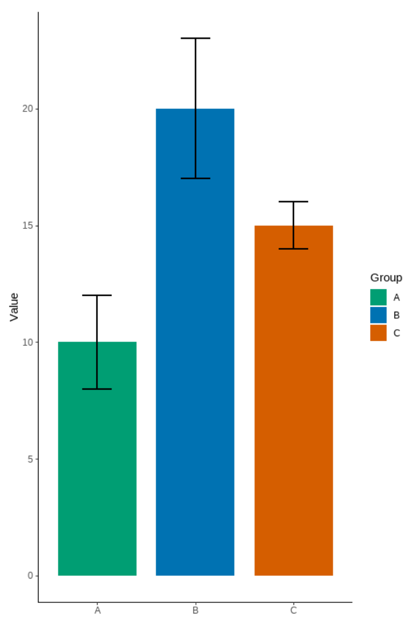


Figure 10

When conducting the designs for the charts and trying to expand our resources with R studio, some design ideas was not fit into the analysis. The figure 10 chart was a design which had come to our knowledge that had some promising impact on how the result could have been shown, but for this research, this design was eventually deleted.

6. Reference

Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. *International Journal of Advanced Computer Science and Applications*, *8*(6).

Alhuzali, H., Zhang, T., & Ananiadou, S. (2022). A comparative geolocation and text mining analysis of emotions and topics during the COVID-19 Pandemic in the UK. *Journal of Medical Internet Research*, *24*(10), N.PAG. https://doi.org/10.2196/40323

Ansari, M. T. J., & Khan, N. A. (2021). Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. *Electronic Journal of General Medicine*, *18*(6). https://doi.org/10.29333/ejgm/11316

Balli, C., Guzel, M. S., Bostanci, E., & Mishra, A. (2022). Sentimental analysis of Twitter users from Turkish content with natural language processing. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/2455160

Chandrasekaran, R., Mehta, V., Valkunde, T., & Moustakas, E. (2020). Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study. *Journal of medical Internet research*, *22*(10), e22624. https://www.jmir.org/2020/10/e22624/'

De Groot, A. M., & Hagoort, P. (Eds.). (2017). *Research methods in psycholinguistics and the neurobiology of language: A practical guide*. John Wiley & Sons.

FHM (n.d.). FHM https://fhm.se/. Retrieved from https://fhm.se/

Garg, N., & Sharma, K. (2022). Text pre-processing of multilingual for sentiment analysis based on social network data. *International Journal of Electrical & Computer Engineering (2088-8708)*, *12*(1). https://www.academia.edu/download/75679121/81_25456_EMr_15Jul_25Mar_NK.pdf

Hauthal, E., Burghardt, D., & Dunkel, A. (2019). Analyzing and Visualizing Emotional Reactions Expressed by Emojis in Location-Based Social Media. *ISPRS International Journal of Geo-Information*, *8*(3), 113. https://doi.org/10.3390/ijgi8030113

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* *https://doi.org/10.1609/icwsm.v8i1.14550*

Janetzko, D. (2017). The role of APIs in data sampling from social media. *The Sage handbook of social media research methods*, 146-160.

Lamsal, R. (2021). Design and analysis of a large-scale COVID-19 tweets dataset. *applied intelligence*, *51*, 2790-2804. https://doi.org/10.1007/s10489-020-02029-z

LeCompte, T., & Chen, J. (2017, December). Sentiment analysis of tweets including emoji data. In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 793-798). IEEE. https://doi.org/10.1109/CSCI.2017.137

Lenstierna, E. (2019). Motivation on Twitter: A qualitative study examining motivational factors for political participation in# svpol.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1.

Melton, C. A., White, B. M., Davis, R. L., Bednarczyk, R. A., & Shaban-Nejad, A. (2022). Fine-Tuned Sentiment Analysis of COVID-19 Vaccine Related Social Media Data: A Comparative Study. *Journal of Medical Internet Research*, *24*(10), N.PAG. https://doi.org/10.2196/40408

Mir, A. A., Rathinam, S., & Gul, S. (2022). Public perception of COVID-19 vaccines from the digital footprints left on Twitter: analyzing positive, neutral and negative sentiments of Twitterati. *Library Hi Tech*, *40*(2), 340–356. https://doi.org/10.1108/LHT-08-2021-0261

The SAGE Handbook of Social Media Research Methods. (2022). SAGE Publications, Ltd. https://doi.org/10.4135/9781529782943

Quinn, E. A., & Monroe, B. L. (2020). sentimentr: An R package for sentiment analysis of text. Journal of Open Source Software, 5(55), 2509. https://doi.org/10.21105/joss.02509

Tidyverse. (n.d.). Tidyverse: Easily Install and Load the 'Tidyverse'. Retrieved from https://tidyverse.org/

Townsend, L., & Wallace, C. (2016). Social media research: A guide to ethics. *University of Aberdeen*, *1*(16). https://www.bolton.ac.uk/assets/Uploads/Social-media-ethics-study-Aberdeen-2018.pdf

Twitter Developers (n.d.). Twitter Developer Platform [Website]. Retrieved from https://developer.twitter.com/en

Rajput, N. K., Grover, B. A., & Rathi, V. K. (2020). Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. *arXiv preprint arXiv:2004.03925*. https://doi.org/10.48550/arXiv.2004.03925

Research group bergman - The AMP-group. (2023, September 2). Karolinska Institutet - ett medicinskt universitet. https://ki.se/en/labmed/research-group-bergman-the-amp-group

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political expression and action on social media: Exploring the relationship between lower-and higher-threshold political activities among Twitter users in Italy. *Journal of Computer-Mediated Communication*, *20*(2), 221-239.

Vohra, A., & Garg, R. (2023). Deep learning based sentiment analysis of public perception of working from home through tweets. *Journal of Intelligent Information Systems*, *60*(1), 255–274. https://doi.org/10.1007/s10844-022-00736-2

Xu, Q. A., Chang, V., & Jayne, C. (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*, 100073. https://doi.org/10.1016/j.dajour.2022.100073

Yurtoz, C. U., & Parlak, I. B. (2019). Measuring the effects of emojis on Turkish context in sentiment analysis. *2019 7th International Symposium on Digital Forensics and Security (ISDFS), Digital Forensics and Security (ISDFS), 2019 7th International Symposium On*, 1–6. https://doi-org.bibproxy.kau.se/10.1109/ISDFS.2019.8757554

7. Appendix

```
library(NLP)
library(syuzhet)
library(wordcloud)
library(wordcloud2)
library(tidyverse)
library(tm)
library(twitteR)
library(sentimentr)
library(dplyr)
library(ggplot2)
library(stringr)
library(RColorBrewer)
library(jsonlite)
library(emojifont)

tweets <- read.csv("C:/Users/alexa/Downloads/Covid.csv")
View(tweets)
tweets_filtered <- tweets %>%
  filter(!str_detect(text, "[åäö]"))

# Conduct sentiment analysis on the tweets using the sentimentr package
sentiment_analysis <- sentiment(tweets_filtered$text)

nrow(tweets_filtered)
nrow(sentiment_analysis$sentiment)
any(is.na(tweets_filtered$text))
any(nchar(tweets_filtered$text) == 0)
```

```
# Add the sentiment scores to the tweets data frame
tweets_filtered$sentiment <- sentiment_analysis$sentiment

# Summarize the sentiment scores
sentiment_summary <- tweets_filtered %>%
  group_by(sentiment) %>%
  summarize(count = n())

sentiment_summary

View(sentiment_summary)

# Combine all the tweets into a single character vector
all_tweets <- paste(tweets$text, collapse = " ")

# Create a wordcloud
wordcloud(all_tweets, max.words = 100)

tweets <- read.csv("C:/Users/alexa/Downloads/Covid.csv")

sentiment_analysis <- get_sentiment(tweets$text)

tweets$sentiment <- sentiment_analysis

filtered_tweets <- tweets %>%
  filter(!str_detect(text, "swedish word"))

filtered_sentiment_summary <- filtered_tweets %>%
  group_by(sentiment) %>%
  summarize(count = n())

filtered_sentiment_summary

view(filtered_sentiment_summary)
```

```
# Read in the CSV file
tweets <- read.csv("C:/Users/alexa/Downloads/Covid.csv")

# Load the Swedish stop words
swedish_stop_words <- stopwords("sv")

# Tokenize the text column
tweets_tokenized <- tweets %>%
  unnest_tokens(word, text)

tweets_filtered <- tweets_filtered %>%
  mutate(id = row_number())

# Filter out the Swedish stop words
tweets_filtered <- tweets_tokenized %>%
  filter(!word %in% swedish_stop_words)

# Group the filtered data by the tweet ID and concatenate the words back into sentences
# Filter out specific words
tweets_filtered <- tweets_filtered %>%
  filter(!word %in% c("covid", "fick", "sen", "att", "inget", "fått", "funkar", "tycker", "visar", "låter", "sen",
"just", "precis", "annat", "svenskar", "2", "st", "sverige", "mindre", "ser", "https", "göra", "också", "många",
"000", "t.co", "pandemic", "hellre", "igen", "vill", "ska", "låta", "mer", "stämmer", "vet", "får", "bör", "abf",
"covid", "19", "vaccin", "hela", "låt", "året", "år", "2022", "2023", "tre", "första", "dit", "år", "aldrig", "total",
"sa", "ja", "nej", "innan", "hittade", "ner", "måste", "alltså", "fram", "enligt", "ta", "åka", "alltid", "svar", "går",
"tittat", "ny", "tills", "ge", "bara", "andra", "alltså", "inga", "heller", "gissa", "kalla", "gapa", "istället", "sitt",
"covid-19", "säger"))


# View the final filtered data
View(tweets_filtered)
# View the final filtered data
View(tweets_filtered)

write.csv(tweets_filtered, "tweets_filtered.csv")
```

```
tweets_filtered <- read.csv("tweets_filtered.csv")

# Create a word frequency table
word_freq <- table(tweets_filtered$word)

# Generate the wordcloud
wordcloud(names(word_freq), freq = word_freq, min.freq = 5, random.order = FALSE)

# Create a word frequency table
word_freq <- as.data.frame(table(tweets_filtered$word))

# Rename the columns of the word frequency table
names(word_freq) <- c("Word", "Frequency")

# Write the word frequency table to a CSV file
write.csv(word_freq, "tweets_word_freq.csv", row.names = FALSE)

# Read in the word frequency table from CSV file
tweets_word_freq <- read.csv("tweets_word_freq.csv")

# Read the CSV file EMOJIS
tweets <- read.csv("C:/Users/alexa/Downloads/Covid.csv")

# Define the emojis to remove
remove_emojis <- c("å", "ä", "ö", "Å", "Ä", "Ö", "ć", "é", "🇦", "🇪",
                   "🇮", "🇲", "🇸", "🇹", "🇺", "🫠", "😕", "🫧", "🛟",
                   "§", "…", "€", "≠", "”", "“", "’", "´", "‼", "—", "–")

# Extract the emojis from the tweets
emojis <- sapply(tweets$text, function(x) {
  str_extract_all(x, "[^[:ascii:]]")
})

# Convert the list of emojis to a vector
```

```
emojis <- unlist(emojis)

# Remove the specified emojis
emojis_filtered <- emojis[!emojis %in% remove_emojis]

# Remove the first 5 rows from the filtered emojis
emojis_filtered <- emojis_filtered[-(1:5)]

# Summarize filtered emojis
emojis_summary <- table(emojis_filtered)

# Remove the first 5 rows the emojis_summary table
emojis_summary <- emojis_summary[-(1:5)]

view(emojis_summary)

# Create a data frame from the emojis_summary table
emojis_df <- data.frame(emoji = names(emojis_summary), count = as.numeric(emojis_summary))


# Create a scatter plot of the emojis_df data frame if it “works”
ggplot(emojis_df, aes(x = emoji, y = count)) +
  geom_point() +
  ggtitle("Emoji Frequency Scatter Plot") +
  xlab("Frequency") +
  ylab("Emoji")

# Create a data frame from the emojis_summary table
emojis_df <- data.frame(emoji = names(emojis_summary), count = as.numeric(emojis_summary))

# Count the number of emojis within each message
emojis_df$emoji_count <- sum(as.numeric(emojis_df$count), na.rm = TRUE)


# Create a scatter plot of the emojis_df data frame
```

```
ggplot(emojis_df, aes(x = emoji_count, y = count)) +
  geom_point() +
  ggtitle("Emoji Frequency Scatter Plot") +
  xlab("Number of Emojis within Message") +
  ylab("Number of Reactions")


ggplot(sentiment_summary, aes(x = sentiment, fill = ..count..)) +
  geom_histogram(binwidth = 0.1) +
  scale_fill_gradient(low = "blue", high = "red") +
  labs(x = "Sentiment Score", y = "Frequency", fill = "Frequency") +
  theme_classic()

# Plot the sentiment summary as a bar chart change the sentiment_summary for other data
ggplot(data = sentiment_summary, aes(x = sentiment, y = count, fill = sentiment)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(limits = c(-1.2, 1.2), expand = c(0, 0)) +
  scale_fill_gradient(low = "red", high = "green") +
  labs(x = "Sentiment Score", y = "Count", fill = "Sentiment Score") +
  theme_minimal()

ggplot(data = sentiment_summary, aes(x = count, y = sentiment, fill = sentiment)) +
  geom_bar(stat = "identity", width = 1, color = "black", size = 0.2) +
  scale_y_continuous(limits = c(-1, 1), expand = c(0, 0)) +
  scale_x_continuous(limits = c(0, 15), expand = c(0, 0)) +
  scale_fill_gradient(low = "red", high = "green") +
  labs(x = "Count", y = "Sentiment Score: Positive & Negative", fill = "VADER Scale") +
  theme_minimal() +
  coord_flip()

# Create the word cloud
ggplot(tweets_text, aes(x = 1, y = 1, size = n, color = n, label = word)) +
  geom_text_wordcloud2(fontSize = function(n) n * 2) +
  scale_size_gradient(low = "white", high = "red") +
  scale_color_gradient(low = "white", high = "red") +
```

```
  ggtitle("Word Cloud of Covid Tweets") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        plot.background = element_blank())

# Plot the sentiment summary as a bar chart
ggplot(data = sentiment_summary, aes(x = sentiment, y = count, fill = sentiment)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(limits = c(-1, 1), expand = c(0, 0)) +
  scale_fill_gradient(low = "red", high = "green") +
  labs(x = "Sentiment Score", y = "Count", fill = "Sentiment Score") +
  theme_minimal()

# Plot the emojis_summary_df data frame
ggplot(data = emojis_summary_df, aes(x = reorder(emojis_filtered, -Freq), y = Freq)) +
  geom_bar(stat = "identity", fill = "black", color = "black") +
  xlab("Emojis") +
  ylab("Frequency") +
  ggtitle("Frequency of Emojis in Tweets") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))


ggplot(data = emojis_summary_df, aes(x = reorder(emojis_filtered, -Freq), y = Freq)) +
  geom_line(color = "black") +
  xlab("Emojis") +
  ylab("Frequency") +
  ggtitle("Frequency of Emojis in Tweets") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))

# Plot the emojis_summary_df data frame
ggplot(data = emojis_summary_df, aes(x = reorder(emojis_filtered, -Freq), y = Freq)) +
  geom_bar(stat = "identity", fill = "Pink", color = "black") +
  xlab("Emojis") +
```

```
  ylab("Frequency") +
  ggtitle("Frequency of Emojis in Tweets") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))



ggplot(sentiment_summary_filtered, aes(x = sentiment, y = count))+
  geom_bar(stat = "identity")+
  labs(x = "sentiment scores", y = "Count")+
  ggtitle("Sentiment Summary of Emojis (negative and positive sentiment)")

ggplot(sentiment_summary_filtered, aes(x = sentiment, y = count, fill = sentiment)) +
  geom_bar(stat = "identity", color = "black", width = 0.6, orientation = "v") +
  labs(x = "Sentiment Score", y = "Count") +
  ggtitle("Sentiment Summary of Emojis (Negative & Positive)") +
  scale_fill_gradient2(low = "blue", high = "green", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name = "Sentiment Score") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
        axis.text.y = element_text(size = 14),
        axis.title.y = element_text(size = 14, face = "bold"),
        axis.text.x = element_text(size = 14),
        axis.title.x = element_text(size = 14, face = "bold"),
        legend.title = element_text(size = 14, face = "bold"),
        legend.text = element_text(size = 14),
        panel.grid.major.x = element_line(color = "gray", size = 0.5),
        panel.grid.minor.x = element_blank())

ggplot(sentiment_summary_filtered, aes(x = sentiment, y = count, fill = sentiment)) +
  geom_bar(stat = "identity", color = "black", width = 0.8, orientation = "v") +
  labs(x = "Sentiment Score", y = "Count") +
  ggtitle("Sentiment Summary of Emojis (Negative & Positive)") +
  scale_fill_gradient2(low = "blue", high = "green", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name = "Sentiment Score") +
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
        axis.text.y = element_text(size = 14),
        axis.title.y = element_text(size = 14, face = "bold"),
        axis.text.x = element_text(size = 14),
        axis.title.x = element_text(size = 14, face = "bold"),
        legend.title = element_text(size = 14, face = "bold"),
        legend.text = element_text(size = 14),
        panel.grid.major.x = element_line(color = "gray", size = 0.5),
        panel.grid.minor.x = element_blank()) +

  ggplot(sentiment_summary_filtered, aes(x = factor(sentiment), y = count, fill = sentiment)) +
  geom_bar(stat = "identity", color = "black", width = 0.2, orientation = "v") +
  labs(x = "Sentiment Score", y = "Count") +
  ggtitle("Sentiment Summary of Emojis (Negative & Positive)") +
  scale_fill_gradient2(low = "gray", high = "gray", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name = "Sentiment Score") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 25, face = "bold"),
        axis.text.y = element_text(size = 10),
        axis.title.y = element_text(size = 16, face = "bold"),
        axis.text.x = element_text(size = 6.35),
        axis.title.x = element_text(size = 16, face = "bold"),
        legend.title = element_text(size = 16, face = "bold"),
        legend.text = element_text(size = 12),
        panel.grid.major.x = element_line(color = "Black", size = 0.7),
        panel.grid.minor.x = element_blank())

# Plot the deviation bar plot
ggplot(data, aes(group, value, fill = group)) +
  geom_bar(stat = "identity", width = 0.8) +
  geom_errorbar(aes(ymin = value - deviation, ymax = value + deviation),
                width = 0.3, size = 1) +
  scale_fill_manual(values = c("#009E73", "#0072B2", "#D55E00")) +
```

```
  labs(x = "Group", y = "Value", fill = "Group") +
  theme_classic()

ggplot(sentiment_summary_filtered, aes(x = sentiment, y = count, fill = sentiment)) +
  geom_bar(stat = "identity", color = "black", width = 0.8, orientation = "v") +
  labs(x = "Sentiment Score", y = "Count") +
  ggtitle("Sentiment Summary of Emojis") +
  scale_fill_gradient2(low = "blue", high = "green", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name = "Sentiment Score") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
        axis.text.y = element_text(size = 14),
        axis.title.y = element_text(size = 14, face = "bold"),
        axis.text.x = element_text(size = 14),
        axis.title.x = element_text(size = 14, face = "bold"),
        legend.title = element_text(size = 14, face = "bold"),
        legend.text = element_text(size = 14),
        panel.grid.major.x = element_line(color = "gray", size = 0.5),
        panel.grid.minor.x = element_blank()) +
  coord_flip()
```