# Comparing popular topics related to Covid-19 in Latin America and Europe

Pranav Chobdar
p.chobdar@student.utwente.nl
University of Twente
Enschede, Netherlands

## ABSTRACT

The coronavirus pandemic ravaged the world with millions of deaths, lockdowns, etc. The pandemic has led to people all around the globe using social media more often to voice their opinions, especially on platforms such as Twitter. The effects and sentiments regarding the coronavirus pandemic have largely varied across different continents. The question remains, how did people on different continents react to the virus? Moreover, whether there are any differences between the sentiments of people from various continents about COVID-19. In this research, the Netherlands and Mexico were chosen as countries to conduct the research. The Netherlands and Mexico are culturally, politically, and socio-economically very different countries in Europe and Latin America, respectively. This paper makes two contributions. Firstly, popular topics related to COVID-19 are found in Mexico and the Netherlands. Lastly, the population's sentiment related to the popular topics is compared. Data from Twitter in the form of tweets is collected from Mexico and the Netherlands during COVID-19. Clustering algorithms are used to perform an analysis of the collected data. Sentiment analysis on the popular topics related to COVID-19 is performed to find sentiment among the populace about the topics. The results aim to compare popular topics about COVID-19 in Latin America and Europe and the sentiments underneath the topics. Overall the research will provide a better understanding of the popular topics and people's sentiments across different continents during the coronavirus pandemic.

## KEYWORDS

Covid-19, Corona , Netherlands , Mexico , Clustering Algorithms , k means , Sentiment Analysis , NRC Lexicon , Twitter Analysis

**ACM Reference Format:**
Pranav Chobdar. 2022. Comparing popular topics related to Covid-19 in Latin America and Europe . In *Proceedings of July 08-07-2022 (TScIT37).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The coronavirus pandemic (Covid-19) has been the most widespread virus pandemic since the beginning of the 20th Century. The virus is responsible for over 400 million cases, and over 6 million deaths [23]. The virus quickly spreads through droplets in the air and has reached an R number of 3. The R number/ Reproduction number indicates how many other people can be infected by a person having the virus [33].

Globally, governments have imposed harsh restrictions and measurements to curb the virus's spread. Measurements such as lockdowns, working from home, and general isolation profoundly impact people's psyche and mental well-being. Therefore, lack of social mobility eventually led to mental health decline and more social media usage amongst the general population [8].

Statements and opinions posted on prominent social media platforms give an insight into the masses' thoughts, emotions, and feelings, thus providing a great tool to analyze sentiments. Twitter is one of the largest social media platforms where people voice their opinions about breaking news and current topics. G7 world leaders have approximately 85 million followers, and 500 million visitors log into Twitter every month [26].

Empirical studies show that the spread of coronavirus has been unequal amongst different countries and socio-economic classes [31]. Another empirical study shows different media narratives across different countries during various peaks of the pandemic around the world [22]. These studies show the impact of the coronavirus on different socio-economic classes and media narratives in different countries. However, research has been conducted on the impact of coronavirus in various countries and socio-economic classes. There is a lack of information on how different populations across different continents perceive the coronavirus pandemic. There is not enough information available about popular topics related to COVID-19 amongst the general populace across Latin America and Europe in a certain time range. This research paper aims to cover the gap caused by the lack of information on popular topics related to COVID-19 amongst the general population in Latin America and Europe.

For this study, two countries in Latin America (Mexico) and Europe (Netherlands) have been chosen. According to Hofstede's chart [13] [19], Netherlands and Mexico have distinctive cultures. Furthermore, there are significant differences between the Netherlands and Mexico regarding press freedom [10] and human development (UNDP 2022). Hence, such large differences might provide an interesting perspective on both sides. The WHO declared a coronavirus pandemic across the globe on 30th March 2020 [2]. This research paper aims to answer the following questions:-

*TScIT37, Enschede,Netherlands,*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

**Q What are the main differences in popular topics related to Covid-19 between Mexico and the Netherlands from 20$^{\text{th}}$ March 2020 to 20$^{\text{th}}$ May 2020?**

**Q What are the main differences regarding sentiments about popular topics related to Covid- 19, between Mexico and the Netherlands from 20th March 2020 to 20 May 2020?**

In order to answer these research questions, topic modeling and sentiment analysis are performed. Data is collected from Twitter in the form of 'tweets,' the text posted by the user on the platform. The data is cleaned, stemmed, and processed into features. Furthermore, excess features are reduced. Finally, clustering algorithms are applied for topic modeling and libraries for sentiment analysis.

The main contributions of this research paper are :-

- To understand and compare popular topics related to COVID-19 in the Netherlands and Mexico. This research will provide an understanding of similarities and differences in trending topics during COVID-19 in these two countries. This will give a better insight into the thoughts and opinions of the population in Mexico and the Netherlands.
- Understand sentiments behind COVID-19 topics in the Netherlands and Mexico. The data in this research will provide a better understanding of the effects of the pandemic on emotions, feelings, and sentiments in these countries.

The remainder of this research paper is organized as follows: The second chapter discusses the literature review, which discusses previous work done in clustering algorithms and sentiment analysis. The third chapter delves deeper into the research methodology, data processing techniques, the K-means algorithm, and NRC sentiment analysis. The fourth chapter is about observations and analysis of the research conducted. Finally, the fifth chapter discusses the results and overall discussion of the research.

## 2 LITERATURE REVIEW

In the past, many researchers have used clustering algorithms and sentiment analysis to research popular topics and the sentiments associated with these topics.

A study by [14] discovered popular COVID-19-related topics worldwide in their research.

Researchers collected data related to COVID-19 worldwide using the Twitter API. The results showed nine clusters of different topics with the highest score of 83.25% positive and 16.75% negative sentiments. Furthermore, the results explored different topics in clusters and visualization of clusters. First, the data was collected, cleaned, filtered, and processed. The terms were found using natural language algorithms. Once the terms were found, they were reduced to fit the size. The features' size was reduced using SVD (Singular Value Decomposition). Once the features were reduced to the most prominent features, clustering algorithms such as the k-means algorithm were applied to these features to find the most popular topics. The data was visualized using t-SNE (t-Distributed Stochastic Neighbor Embedding). Finally, sentiment analysis was performed on popular topics to find overall sentiment related to these topics. Software related to sentiment analysis, such as TextBloB, was used to find underneath sentiments. The research, however, lacks the usage of various other clustering algorithms. Moreover, in this research, we compare COVID-19 popular topics and sentiments in two different countries. Comparison between two countries gives in-depth view and helps compare sentiments and popular topics in different cultures and languages.

Another research study by [15] compared different clustering algorithms on the Covid-19 data set to demonstrate the different accuracy of clustering algorithms. The research found k-means algorithms as the best clustering algorithm. The research stated that the k-means algorithm helps find the best quality clusters with less computational time.

Lastly, a similar research study conducted by [9], compares topic modeling and performs a sentiment analysis between the USA and Brazil. The findings compared English and Portuguese tweets. The findings showed that the overall sentiment in English and Portuguese was similar; in this case, it was negative. Moreover, seven of ten popular topics were similar in Portuguese and English. The study uses the Gibbs Sampling algorithm for Dirichlet Multinomial Mixture (GSDMM) for topic modeling [9]. GSDMM is a generative process to build documents where terms/words are drawn from a probability distribution, with each probability distribution representing a topic. However, another study [30] states that the k-means algorithm with TF-IDF is best for comparing cluster analysis results to an externally known provided class label. Although the study by [9] and this research paper are similar in comparing topic modeling and sentiment analysis between two countries. This study compares Spanish and English tweets

This study aims to use TF-IDF (Term Frequency-Inverse Document Frequency) to find the most popular terms. Excess features/terms will be removed using Singular Value Dimension (SVD) to extract the most relevant features. Furthermore, the K-means clustering algorithm will be used to perform the topic analysis. Clusters will be visualized using t-SNE (t-Distributed Stochastic Neighbor Embedding). Finally, sentiment analysis will be performed using software such as NRC Lexicon, which describes the sentiment in 8 different emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust).

## 3 DATA ANALYSIS FRAMEWORK

This section will talk about the data analytical framework used in this study. Using various data analysis techniques on Twitter data gives a more in-depth view of the underlying topics and sentiments. The process is as follows:

### 3.1 Data Collection

The data collected for this study was done in following steps:-

- The IEEE dataset of COVID-19 tweets from [17] was used as the source to extract tweets related to the corona pandemic. The dataset contains about 2 billion tweets from all around the globe. The tweets have been translated into English. Tweets ranging from March $20^{th}$ 2020 till May $20^{th}$2020 were downloaded from the IEEE dataset. Due to Twitter's policy, tweets were provided as tweet ids, which were further processed.

- The Twitter API, along with the software library of twarc, were used to hydrate the tweet ids. Twarc is a Python library used to collect Twitter data through the Twitter API. Twarc software was used to hydrate the tweet ids. Hydration is a process where the text from a tweet is extracted through a given tweet id.

## 3.2 Data Filtration and Processing

The collected Tweets contain irrelevant data, which gives inaccurate results. Therefore, tweets undergo a filtration process to remove the noise.

*3.2.1 Data Filtration.* Tweets contain many different words and symbols, many of which have no meaning in the context of the information they contain. Symbols such as hashtags (), mentions (@), URLs (e.g. https:///abc.org) , digits and special characters, empty tweets and duplicate tweets are removed from the tweets. Punctuations (e.g.,.) and abbreviations (e.g greeeaat for great) are also removed. Python code is written to keep track of the criteria above and filter out unnecessary noise from tweets [1]

*3.2.2 Data Processing.* Stopwords are the most frequent words used in a language. In English, "a" and "at" are examples, while in Spanish, "un," "sobre," and "todo" are examples. Stopwords do not add value to the context and thus need to be removed [28]. Stopwords are removed using the help of NLTK's stopword list, which is available in both English and Spanish.

Stemming is a normalization process where the prefix and the suffix are removed from certain words to give them the same base [14]. For example, eating, eats, eaten are stemmed into "eat." Stemming helps reduce the exact words' overall volume, thus increasing the overall accuracy of the results. A Python library called snowballStemmer is used, which contains a stemmer for both English and Spanish languages.

## 3.3 Tokenization

Once the data is filtered and processed, a tokenization technique is applied to the tweets. Tokenization segments the tweets into different words, which are referred to as tokens/features. Features are used in clustering algorithms for topic modeling. Features can be extracted from a tweet in different ways, such as n-grams representation, word frequency, Latent Segment Allocation, or TF-IDF (Term Frequency-Inverse Document Frequency). This study used the TF-IDF technique to extract features from the tweets. The TF-IDF technique is used to find the importance of a word in a set of documents. TF-IDF uses a certain word's term frequency (TF), which occurs in a set of documents. However, as stated in [14], term frequency in itself is inadequate to give accurate information about the importance of a word in a list of documents. In this research, the TF-IDF Vectorizer library from SK-Learn is implemented.

## 3.4 Reducing Excess Features

Once the features are collected, irrelevant or excess features must be removed. Too many features can cause high dimensionality. High dimensionality is a common problem in data analytics studies when mining data. High dimensionality occurs when the number of features exceeds the number of observations. High dimensionality can cause computation problems as data becomes more sparse [7]. Feature reduction techniques such as SVD (Singular Value Dimension), PCA (Principle Component Analysis), Linear Discriminant Analysis (LDA), and Non-negative matrix factorization (NMF) are used to reduce the number of features. This study uses the SVD (Singular Value Dimension) technique to reduce the number of features. According to [32], SVD reduces the number of features and is less computationally expensive than PCA. As a result, SVD is a good choice for feature reduction [25]. Truncated SVD library from sk-learn.decompisition is used in this study for feature reduction." The size of the matrix can be reduced to n components so that only the most relevant components are kept. The explained variance can determine the value of n components. An estimated value for variance should be above 95% [14]

After all these steps in this section, tweets are filtered and processed, and the most relevant features are extracted. The next step is clustering and topic modeling, along with sentiment analysis.

## 3.5 K-Means Clustering Algorithm

K-means is a clustering algorithm that uses a set of n data points in a dimension and an integer k. The algorithm tries to make k centers that contain similar data points, and the algorithm does so by calculating the mean squared distance of each data point closest to its near center [24].

*3.5.1 Benefits and Limitations of K- Means.* This section explores some benefits and limitations of K-Means as a clustering algorithm. This is explained as follows:

Advantages

- K-Means is a fast (linear time complexity), and simple algorithm which can handle large data sets with ease [15].
- The K-Means algorithm guarantees convergence of the data points into clusters. Clusters can be of different shapes, such as circles or elliptical [3].
- According to [15], the K-Means algorithm was used along with different clustering algorithms on the COVID-19 data set. As a result, K-Means provided the most accurate results compared to other clustering algorithms.

Disadvantages

- K-Means requires a specific number of clusters 'k'. For example, algorithms such as the mean-shift algorithm do not require a predefined number of clusters to start the algorithm [29].
- According to [18], K-Means helps reduce inter-cluster variance but does not reduce overall global variance in the dataset.
- If there are many outliers, the sum of square errors would increase, thus leading to inaccurate cluster centers [11].

*3.5.2 Finding K.* K-Means algorithms require an input value of 'k' to run the algorithm. Here, 'k' stands for the number of clusters that are meant to be formed. The Elbow method is the most common method to find the value of 'k.'

The Elbow method runs the K means algorithm for values of k ranging from 1–10. As clusters get calculated, the value of the within-cluster sum of squares decreases. Eventually, the graph of k vs within-cluster sum square takes an elbow shape, as shown below [16]. There is a point in the graph where there is a sudden decrease in the value of k, indicating less distortion. Henceforth, the value of k is chosen to be at that point [6]. The sharp declines gives the graph a curved edged, thus giving it an elbow shape. In figures 1 and 2 represents 'k' vs WCSS (Within cluster square sum) graph for the Netherlands and Mexico,respectively. Y-axis reprents WCSS and X-axis represents value of 'k'.

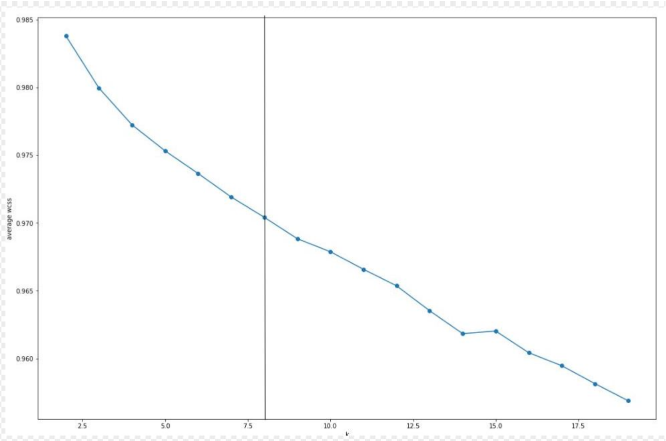
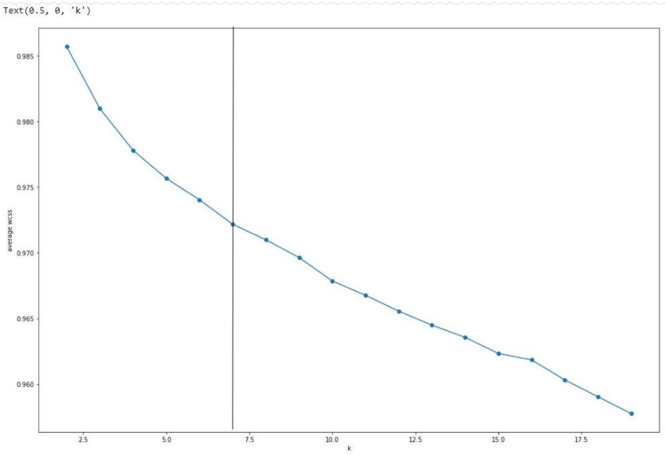


**Figure 1: K Elbow Dutch, k = 8**



```
Text(0.5, 0, 'k')
```

**Figure 2: K Elbow Mexican, k = 7**

### 3.6 Sentiment Analysis : NRC Lexicon

Sentiment analysis is conducted after the k-means algorithm divides the data points into different clusters. Positive, negative, anger, anticipation, joy, anger, trust, sadness, disgust, surprise, and fear are all included in the NRC Lexicon [20]. NRC Lexicon is a sentiment analysis library that has associated emotion categories with each specific word. For example, the word "hate" is associated with sentiments such as fear, anger, and negativity. Furthermore, the NRC Lexicon contains both manually created (joy, anger, fear, etc.) and automatically generated (positive, negative) features [20]. Each emotion and sentiment associated with the word is given a score. NRC Lexicon gives an in-depth analysis of various emotions and sentiments in the data set, especially compared to other sentiment analysis libraries such as TextBlob (only 3 sentiments). The NRC Lexicon python library was used to conduct the sentiment analysis on the dataset.

## 4 EXPERIMENTATION AND OBSERVATIONS

This section will cover the experimentation and observations in the study. This section is divided into 5 parts. The first part will discuss extracting tokens from the given dataset of tweets and reducing features using Singular Value Dimension (SVD). The second part discusses finding a sufficient number of clusters by finding the value of 'k.' The value of 'k' is found using the Elbow method. The third part discusses clustering data along with word clouds for each cluster. Furthermore, the fourth section discusses different categories of topics for both the Netherlands and Mexico. Finally, the fifth section discusses sentiment analysis conducted on the data set.

### 4.1 Feature Engineering

About 15,000 tweets for both the Netherlands and Mexico were analyzed.

The tweets were dated from $20^{th}$ March 2020 to $20^{th}$ May 2020. On average, 2000 tweets per week were selected while constructing the dataset. The tweets were hydrated from the tweet IDs using the python code.

Furthermore, tweets were cleaned and filtered using python code. Finally, the tweets were ready for feature extraction using TF-IDF (Term Frequency-Inverse Document Frequency). About 12000 tokens were extracted from the dataset using TF-IDF.

Once all the features were extracted, feature reduction techniques were applied to extract the most valuable and relevant features. Singular Value Dimension (SVD) was applied for feature reduction. In order to find the optimal number of features, the explained variance of SVD was applied. According to [14], a component size that gives an explained variance of more than 95% is sufficient. The research found that the variance for about 5500 components gave over 95% variance in both the datasets (Netherlands and Mexico).

### 4.2 Clusters and Silhouette Score

K-Means algorithm was applied with k = '8' for the Netherlands and k = '7' for Mexico. The graph here shows clusters in different colors. Each cluster represent set of similar words which are aggregated together by the K-Means algorithm. The graph here was represented using t-SNE software (t-distributed stochastic embedded). Figures 5 and 6 represent cluster graph for the Netherlands and Mexico respectively.

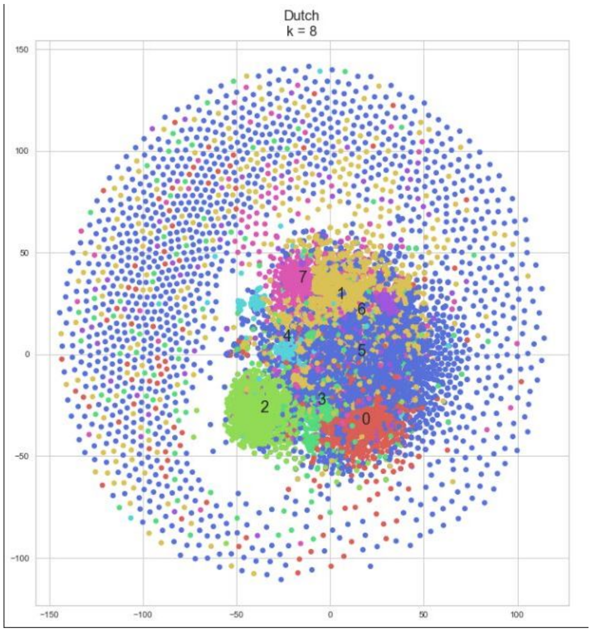


**Figure 3: Dutch Clusters**

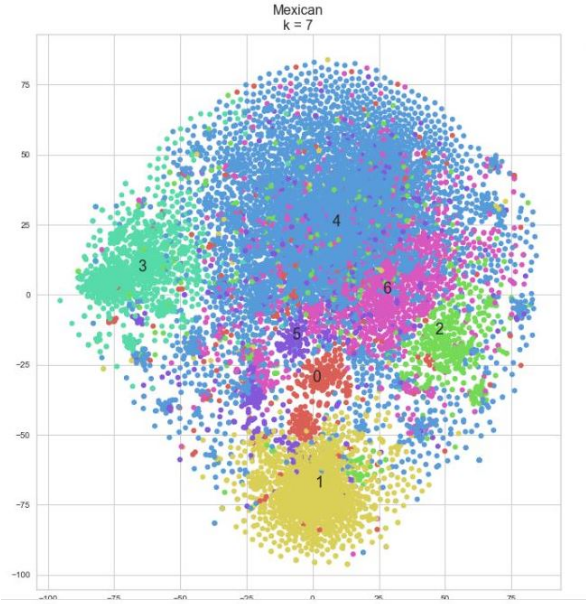


**Figure 4: Mexican Clusters**

*4.2.1 Cluster Quality and Analysis.* As is visible from Figure 3, Cluster 4 has the highest number of data points among the Dutch clusters. In Figure 4, Cluster 5 has the number of data points among Mexican clusters. However, the clusters with most data points had a slightly negative silhouette score (-0.02), indicating that the most significant clusters were not very well clustered. However, the rest of the clusters had a slightly positive score with the average number of data points. Therefore, on average, clusters in Dutch and Mexican graphs were clustered well.

Finally, Dutch and Mexican clusters achieved an overall low silhouette score 0.007. Furthermore, Dutch Clusters had more outliers on average as shown in the graph. The reasoning for low silhouette scores can be explained by [14], which states, "challenges of unstructured short textual data analysis, such as the data representation, weighting scheme, high dimensionality, sparse feature vector, and the word synonymy. Moreover, some challenges are faced by the textual data clustering technique, such as the similarity measures, determining the optimal centers and the huge number of outliers".

## 4.3 Sentiment Analysis

About 15000 tweets were analyzed to understand the underlying sentiments and emotions behind the tweets. The tweets were filtered and processed as discussed in section 3.2 to remove the noise and irrelevant information. Furthermore, the NRC lexicon library was applied to the tweets. NRC Lexicon library results in 8 emotions (fear, anger, sadness, disgust, joy, anticipation, trust, surprise) and 2 polarities (Positive and Negative). Figures 5 and 6 represent the emotion count vs emotion graph for Netherlands and Mexico, respectively. X-axis represents the emotion count. Y axis represents the emotions. From the graphs, there are some similarities and differences that are visible.

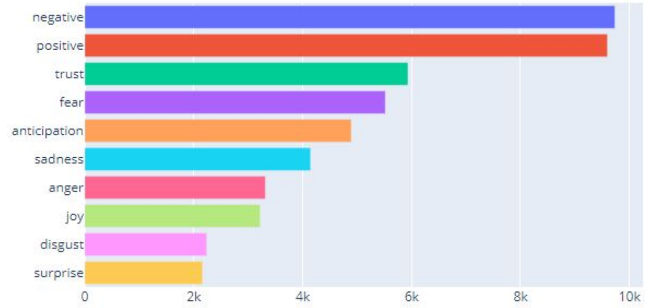


**Figure 5: Dutch Sentiments**

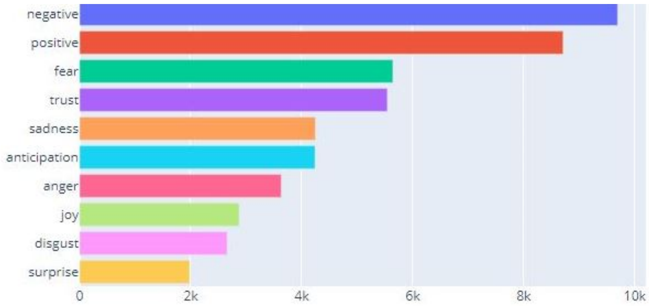


**Figure 6: Mexican Sentiments**

In both Figure 5 and Figure 6, negative emotions seem to have the highest count in both countries, with approximately 10000 emotions count. Trust and fear seem to be high in both countries, with around 6000 emotion counts. There is slightly more anticipation in the Dutch sentiments (5000) compared to the Mexican sentiments (4100). The rest of the emotions, like anger, sadness, joy, disgust, and surprise, seem to be on par in both Dutch and Mexican sentiments.

## 4.4 Word Topics and Classification

The words inside the clusters are represented in the form of word clouds. Each word cloud contains words similar to each other, which represent similar topics. Overall, there are '8' word clouds for the Netherlands and '7' word clouds for Mexico. Figures 7 and 8 represent word clouds for Netherlands and Mexico, respectively.

In order to gain a better understanding of the words in the word clouds, words were classified into certain topics. Therefore, the most frequently occurring words were manually classified into 7 topics. The topics are related to education, health, proliferation care, treatment, politics, economic impact, and daily life. Furthermore, graphs for topic vs. volume of messages were made to understand which topics were most prevalent in both countries.

Firstly, Tables 1 and 2 represent classified topics for Netherlands and Mexico. Figures 7 and 8 represent graphs of the volume of messages vs. classification for the Netherlands and Mexico, respectively. The x-axis represents the volume of messages. The y-axis represents the topics. It is visible that in both the graphs, different topics received extra attention. The tables and graphs are shown below:-

Table 1: Dutch Topics

| Topic | Correlated Words |
|---|---|
| Economic Impact | lockdown,industri,affect,crisis,work |
| Treatments | vaccine,disease,drug,medic,hospital,patient,help |
| Proliferation Care | stay,home,death,flu,care,protect,mask |
| Case Statistics | dutch,pandemic,infect,die,report,outbreak,update,new |
| Politics | china,trump,rutte,UK,johnson,amsterdam,netherland |
| Daily Life | manage,share,live,fight,talk,watch,miss |

Table 2: Mexican Topics

| Topic | Correlated Words |
|---|---|
| Economic Impact | crisis,lockdown,masterstroke,hike,impact,worker |
| Treatments | virus,patient,positive,help,inject,cure,health,treatment |
| Proliferation Care | people,stay,home,report,medic,disinfect,face,quarantin |
| Case Statistics | pandemic,case,death,die,news,grandpa,grandson,million |
| Politics | trump,president,mexico obrador,referendum,imprison,amlo |
| Daily Life | think,want,kiss,whatsapp,train,plane,garlic,utensil,fight,beer |

Figure 7: Dutch Classification

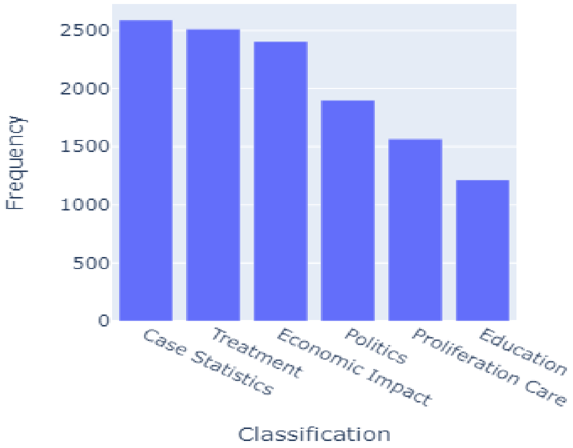
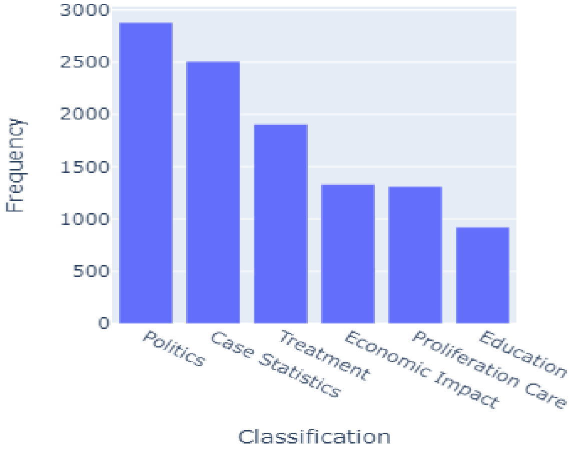


Figure 8: Mexican Classification



From Tables 1 and 2, it is visible that there are some similarities between the words in specific topics. In topics like treatment and proliferation care, topics related to Covid-19 like wearing masks, staying home, and quarantine seem to be a common discussions. Furthermore, both countries seem to be discussing hospitals and patients testing positive due to Covid-19. In Figure 7, it is visible that in the Netherlands, case statistics (2500 words) is the most discussed topic, followed closed by Treatment (2400 words). Economic Impact (2300 words) is also prevalent in the Netherlands. In Mexico (Figure 8), case statistics is a reasonably popular topic (2500 words). Case statistics popularity can be explained by continuous reporting of corona-related tragedies and deaths by various news and media outlets in both countries. However, Politics emerges to be the most popular topic by far in Mexico by a margin (2800 words).

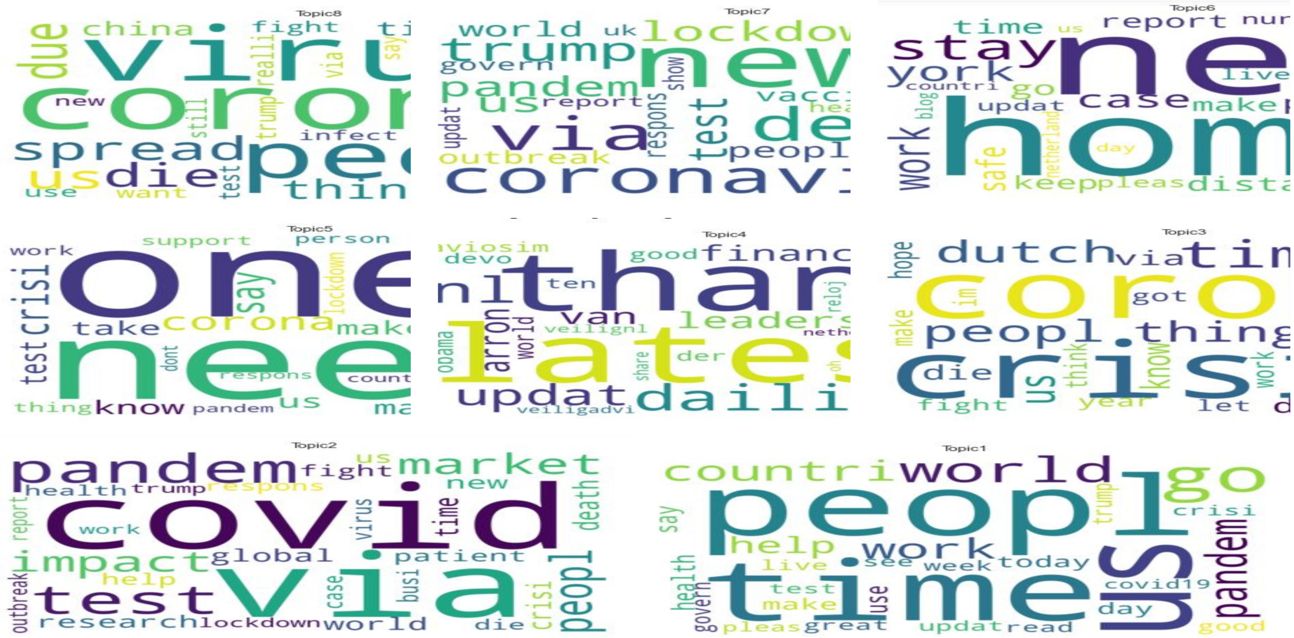
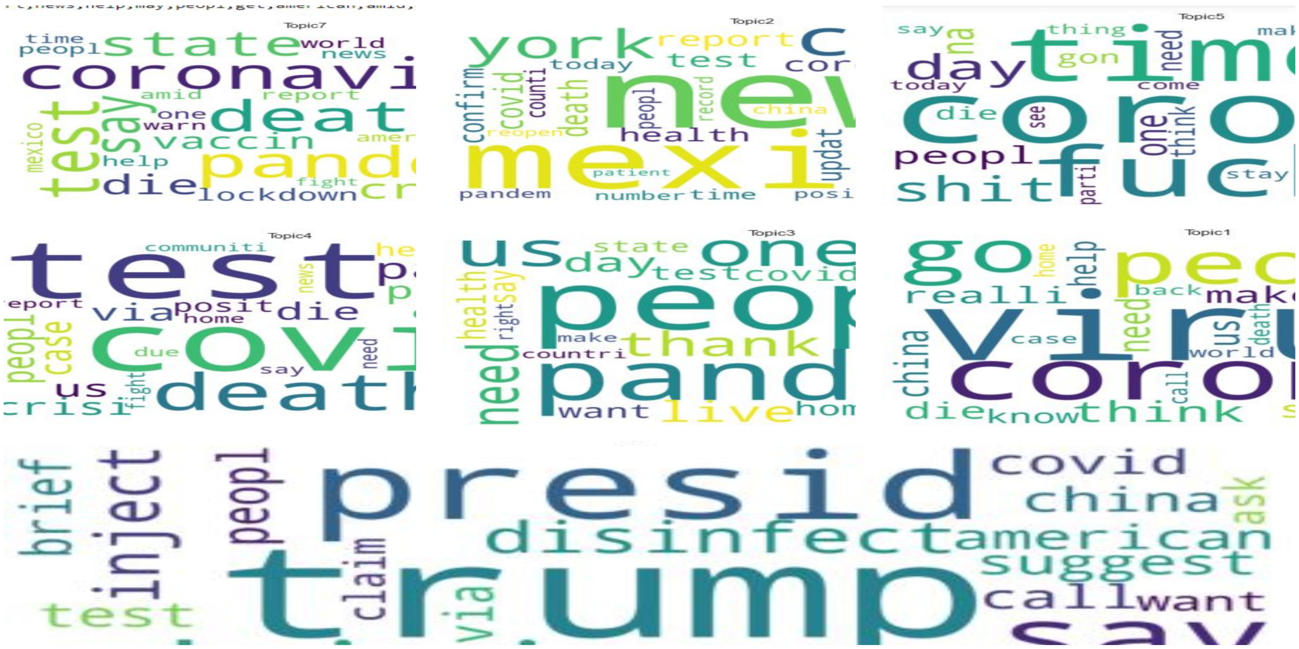


**Figure 9: Dutch WordCloud**

**Figure 10: Mexican WordCloud**



## 5 RESULTS AND DISCUSSION

This section will cover the results of the observations and experiment section in more detail. Firstly, the section discusses potential reasons behind sentiment analysis observations. Lastly, the unit will discuss more results found in the word cloud and topics, highlighting the differences and similarities in popular topics of COVID-19 between the Netherlands and Mexico.

As discussed in section 4.3, Mexican and Dutch sentiments seem negative overall, along with high sentiments for fear and trust. The further analysis gave more insights into tweets associated with fear and trust. Most tweets in both countries related to fear and negative sentiments were related to the coronavirus pandemic. Tweets related to corona deaths, rising cases, seeking help and aid, and complaining about the virus seemed to be the hottest topics. Negative tweets about Donald Trump and his corona-related policies seemed to be a point of criticism in both countries.

In the Netherlands, tweets criticizing the intelligent lockdown were prevalent amongst the tweets. On March 23, 2020, [12], the Dutch government passed the orders for an intelligent lockdown in the Netherlands. Lockdown policies received criticism on Twitter. Some tweets criticized the government for acting similar to a surveillance state. Furthermore, privacy related to corona tracking apps was also discussed. Moreover, some tweets shared concern about the children's mental health due to schools being shut down.

In the Netherlands, tweets also reflected the news about UK's prime minister Boris Johnson. Boris Johnson was hospitalized for coronavirus on April 6th 2020. Furthermore, Boris Johnson's late implementation of lockdown was criticized in the tweets. According to [5], if the lockdown in the UK was implemented two weeks ago, the infection rate could have gone down by 5% resulting in fewer deaths.

The economic impact due to the coronavirus pandemic in both countries was reflected in the tweets. The lockdown policies, which enforced working from home, social distancing, and other factors, resulted in the loss of jobs in both countries [21][12]. In Mexico, tourism makes a big part of the GDP growth [21]. Tourism in Mexico took a hit due to corona policies across the globe, resulting in a loss of tourism. For example, 95% of restaurants in Mexico city were forced to close their business, and more than 140000 rooms had to shut down in Cancun, Mexico. In the Netherlands, the Dutch economy impacted the GDP rate going down in 2020 by -3.8% and GDP going by 1.5% [21]. Furthermore, due to strict lockdown policies, businesses were forced to closed, and public transportation went down.

Criticism about Donald Trump's policies seems to be reflected in the tweets of both countries. According to [27], Donald Trump played down the impact of the coronavirus along with spreading false information since the start of the pandemic [27]. Furthermore, Trump shared false information on Twitter about the effectiveness of hydroxychloroquine and azithromycin against the coronavirus around march 2020 [27].

In Mexico, president Andrés Manuel López Obrador (AMLO) was criticized heavily in the Mexican tweets. The tweets mainly focused on the mishandling of the coronavirus pandemic by the president AMLO. According to [4], the Mexican government underestimated the coronavirus pandemic. Thus the medical staff was unprepared for the oncoming virus. The medical staff was poorly funded, with insufficient equipment to fight the virus. As a result, on 11th May 2020, more than 8000 Mexican medical staff got infected by the coronavirus [4].

## 6 CONCLUSION

As the coronavirus pandemic raged worldwide, the world shifted to prominent social media like Twitter to present their opinions about the current event. Thus, finding and comparing similarities and differences in the coronavirus across different countries becomes increasingly important. This study established some similarities and differences by comparing the coronavirus topic and sentiments in the Netherlands and Mexico.

Firstly, In both countries, negative sentiments along with fear and trust were the most prevalent ones. Moreover, in both countries, tweets related to coronavirus death, cases, and statistics seemed to be the most popular topic. Secondly, Both countries faced similar challenges with economic impact resulting from the coronavirus pandemic, such as lockdowns and unemployment. Thirdly, criticism related to Donald trump's policies and response to the pandemic was common in both countries. In the Netherlands, popular topics revolved around the intelligent lockdown and its impact on society, such as the mental health of kids, surveillance society, economic crisis, UK's Boris Johnson, and his health. However, in Mexico, popular topics revolved around the mishandling and bad leadership by president Andrés Manuel López Obrador (AMLO), deaths of the medical staff due to poorly funded equipment by the government, and unavailable hospitals.

In the future, more research could be done on a weekly basis on the popular topics and sentiments. Weekly analysis can provide a better insight into changing trends and topics per week and provide a more clear understanding of the research. Furthermore, machine learning algorithms, along with different sentiment analyzers, can be used to find the most accurate sentiment analyzer for the document. Lastly, refined techniques for tokenization of textual data can be implemented for more accurate results.

## REFERENCES

[1] Norjihan Abd Ghani and Siti Syahidah Mohamad Kamal. 2015. A sentiment-based filteration and data analysis framework for social media. (2015).

[2] Saira Baloch, Mohsin Ali Baloch, Tianli Zheng, and Xiaofang Pei. 2020. The coronavirus disease 2019 (COVID-19) pandemic. *The Tohoku journal of experimental medicine* 250, 4 (2020), 271–278.

[3] Leon Bottou and Yoshua Bengio. 1994. Convergence properties of the k-means algorithms. *Advances in neural information processing systems* 7 (1994).

[4] Claudia Caldera-Villalobos, Idalia Garza-Veloz, Nadia Martínez-Avila, Iván Delgado-Enciso, Yolanda Ortiz-Castro, Griselda A Cabral-Pacheco, and Margarita L Martinez-Fierro. 2020. The Coronavirus Disease (COVID-19) challenge in Mexico: a critical and forced reflection as individuals and society. *Frontiers in public health* 8 (2020), 337.

[5] Tim Colbourn. 2020. Unlocking UK COVID-19 policy. *The Lancet Public Health* 5, 7 (2020), e362–e363.

[6] Mengyao Cui et al. 2020. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance* 1, 1 (2020), 5–8.

[7] Jianqing Fan and Runze Li. 2006. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133* (2006).

[8] Junling Gao, Pinpin Zheng, Yingnan Jia, Hao Chen, Yimeng Mao, Suhong Chen, Yi Wang, Hua Fu, and Junming Dai. 2020. Mental health problems and social media exposure during COVID-19 outbreak. *Plos one* 15, 4 (2020), e0231924.

[9] Klaifer Garcia and Lilian Berton. 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied soft computing* 101 (2021), 107057.

[10] Shelton A Gunaratne. 2002. Freedom of the press: A world system perspective. *Gazette (Leiden, Netherlands)* 64, 4 (2002), 343–369.

[11] Shalmoli Gupta, Ravi Kumar, Kefu Lu, Benjamin Moseley, and Sergei Vassilvitskii. 2017. Local search methods for k-means with outliers. *Proceedings of the VLDB Endowment* 10, 7 (2017), 757–768.

[12] Lieke Michaela Hoekman, Marlou Marriet Vera Smits, and Xander Koolman. 2020. The Dutch COVID-19 approach: Regional differences in a small country. *Health Policy and Technology* 9, 4 (2020), 613–622.

[13] Geert Hofstede. 2009. Geert Hofstede cultural dimensions. (2009).

[14] Adnan Hussein, Farzana Kabir Ahmad, and Siti Kamaruddin. 2021. Cluster Analysis on Covid-19 Outbreak Sentiments from Twitter Data using K-means Algorithm. *Journal of System and Management Sciences* (12 2021). https://doi.org/10.33168/JSMS.2021.0409

[15] Halat Ahmed Hussein and Adnan Mohsin Abdulazeez. 2021. Covid-19 pandemic datasets based on machine learning clustering algorithms: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology* 18, 4 (2021), 2672–2700.

[16] Trupti M Kodinariya and Prashant R Makwana. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal* 1, 6 (2013), 90–95.

[17] Rabindra Lamsal. 2020. Coronavirus (COVID-19) Tweets Dataset. https://doi.org/10.21227/781w-ef42

[18] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36, 2 (2003), 451–461.

[19] Ludwien Meeuwesen, Atie van den Brink-Muinen, and Geert Hofstede. 2009. Can dimensions of national culture predict cross-national differences in medical communication? *Patient education and counseling* 75, 1 (2009), 58–66.

[20] Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada* 2 (2013), 234.

[21] Carlos Monterrubio. 2022. The informal tourism economy, COVID-19 and socioeconomic vulnerability in Mexico. *Journal of Policy Research in Tourism, Leisure and Events* 14, 1 (2022), 20–34.

[22] Reuben Ng, Ting Yu Joanne Chow, and Wenshu Yang. 2021. News media narratives of Covid-19 across 20 countries: Early global convergence and later regional divergence. *PLoS One* 16, 9 (2021), e0256358.

[23] World Health Organization et al. 2022. COVID-19 weekly epidemiological update, edition 84, 22 March 2022. (2022).

[24] Bhagwati Charan Patel and GR Sinha. 2010. An adaptive K-means clustering algorithm for breast image segmentation. *International Journal of Computer Applications* 10, 4 (2010), 35–38.

[25] Rhonda D Phillips, Layne T Watson, Randolph H Wynne, and Christine E Blinn. 2009. Feature reduction using a singular value decomposition for the iterative guided spectral class rejection hybrid classifier. *ISPRS Journal of Photogrammetry and Remote Sensing* 64, 1 (2009), 107–116.

[26] Sohaib R Rufai and Catey Bunce. 2020. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *Journal of public health* 42, 3 (2020), 510–516.

[27] Paul E Rutledge. 2020. Trump, COVID-19, and the War on Expertise. *The American Review of Public Administration* 50, 6-7 (2020), 505–511.

[28] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. (2014).

[29] Mehak Nigar Shumaila. 2021. A Comparison of K-Means and Mean Shift Algorithms. (2021).

[30] Vivek Kumar Singh, Nisha Tiwari, and Shekhar Garg. 2011. Document clustering using k-means, heuristic k-means and fuzzy c-means. (2011), 297–301.

[31] Mazou Ngou Temgoua, Francky Teddy Endomba, Jan René Nkeck, Gabin Ulrich Kenfack, Joel Noutakdie Tochie, and Mickael Essouma. 2020. Coronavirus disease 2019 (COVID-19) as a multi-systemic disease and its impact in low-and middle-income countries (LMICs). *SN Comprehensive Clinical Medicine* 2, 9 (2020), 1377–1387.

[32] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. 2016. Fast algorithms for robust PCA via gradient descent. *Advances in neural information processing systems* 29 (2016).

[33] Zian Zhuang, Shi Zhao, Qianying Lin, Peihua Cao, Yijun Lou, Lin Yang, Shu Yang, Daihai He, and Li Xiao. 2020. Preliminary estimates of the reproduction number of the coronavirus disease (COVID-19) outbreak in Republic of Korea and Italy by 5 March 2020. *International Journal of Infectious Diseases* 95 (2020), 308–310.