SENTIMENT ANALYSIS OF COVID-19 VACCINE RESPONSES IN MEXICO

Jessica Salinas, Carlos Flores, Hector Ceballos and Francisco Cantu

School of Engineering and Sciences,
Tecnologico de Monterrey, Monterrey, Mexico

ABSTRACT

The amount of information that social networks can shed on a certain topic is exponential compared to conventional methods. As new COVID-19 vaccines are approved by COFEPRIS in Mexico, society is acting differently by showing approval or rejection of some of these vaccines on social networks. Data analytics has opened the possibility to process, explore, and analyze a large amount of information that comes from social networks and evaluate people's sentiments towards a specific topic. In this analysis, we present a Sentiment Analysis of tweets related to COVID-19 vaccines in Mexico. The study involves the exploration of Twitter data to evaluate if there are preferences between the different vaccines available in Mexico and what patterns and behaviors can be observed in the community based on their reactions and opinions. This research will help to provide a first understanding of people's opinions about the available vaccines and how these opinions are built to identify and avoid possible misinformation sources.

KEYWORDS

Twitter, Data Mining, Sentiment Analysis, Machine Learning, COVID-19.

1. INTRODUCTION

Ever since COVID-19 was declared a global pandemic by the World Health Organization (WHO), different institutions and laboratories around the world have started a race against time to develop a vaccine with the highest possible efficacy. After a year of pandemic, the first vaccines are now available, and the vaccination process has started throughout the globe. However, this process has been affected by the opinions of people regarding the different vaccines [1].

Even though clinical studies have shown the efficacy of the vaccines against COVID-19, many people are not convinced of vaccination and have avoided this process. This behavior has become so dangerous, that the World Health Organization has included vaccine hesitancy in its top 10 global health threats in 2019 [2].

However, COVID-19 is not the only pandemic in which this phenomenon has occurred. When the H1N1 virus was first detected in the United States and spread across the world in 2009, Twitter was full of speculations and conspiracy theories about the origin of the virus and the development of vaccines. Given the influence of social networks on people's opinions and the effect on communication, previous studies have made use of Twitter data to provide an analysis of sentiments toward a specific topic or situation. For instance, the tweets developed during the H1N1 epidemic helped in the search for sentiments such as humor, sarcasm, frustration, relief, misinformation, and more [3]. Also, a similar situation was presented in the study by Bessi et al.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 163-178, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111415

[4], where it was found that friends played a major role in the exposure of false information when Ebola appeared at the end of 2013.

Regarding the current pandemic, recent sentiment analyses that have been performed have dealt with topics such as anxiety and panic [5], lack of support for isolation [6], prediction of self-awareness of precautionary procedures, forecasting of stock sectors [7], and public attitude towards the pandemic, among others. With the given data about tweets, data mining techniques can be applied to processed data to discover patterns among data, build networks, form clusters, and perform classification for making predictions that can provide valuable information about people and their behavior towards the current pandemic.

Given the tremendous effect of social networks on people's opinions and behaviors, this research aims to use data mining techniques to identify the sentiments of the Mexican population towards the different COVID-19 vaccines available in Mexico using Twitter data. The results of this study will provide an insight into the general opinion of the Mexican population regarding these vaccines, and the possibility to identify potential misinformation sources and concerns.

The overview of the proposed methodology is show in Figure 1 and its main contributions are:

- To provide a framework through which the general response of the Mexican population towards the different vaccines available in the countrycan be deciphered.
- To report insights and findings within Twitter data about the behavior of the Mexican population regarding the available vaccines in the country.
- To find and report different topics within the COVID-19 vaccine’s conversations of Mexicans on Twitter using novel text mining approaches.
- To identify the main contributors/user profiles on the Mexican population on Twitter regarding the different COVID-19 vaccines using Social Network Analysis.

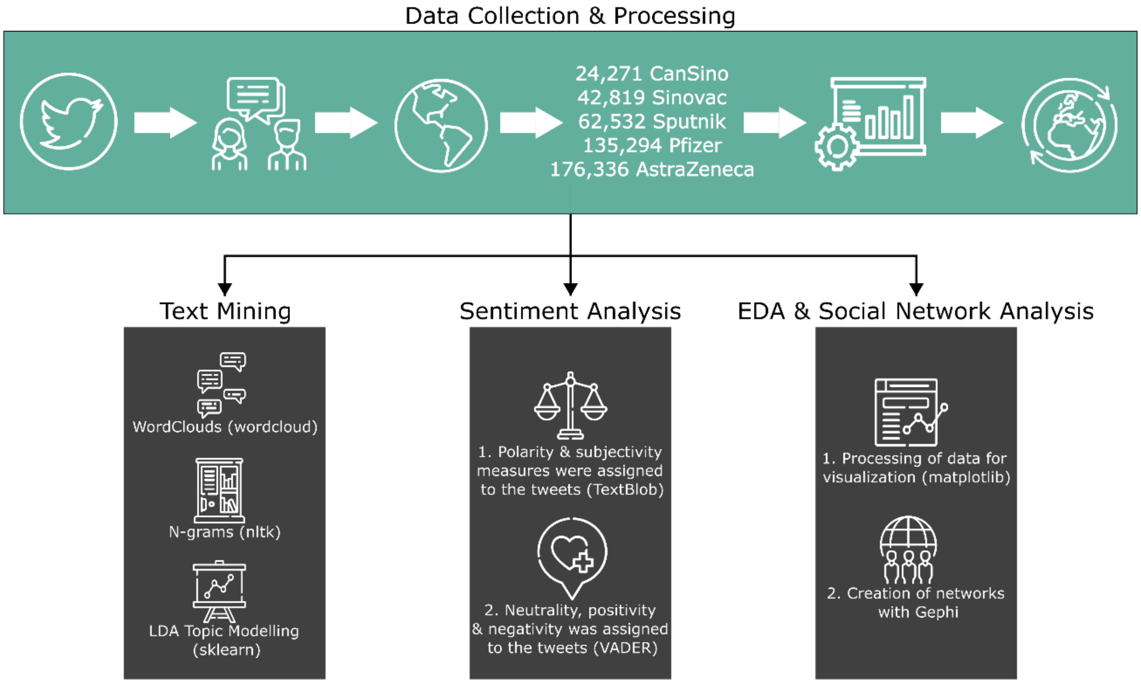
Data Collection & Processing
24,271 CanSino
42,819 Sinovac
62,532 Sputnik
135,294 Pfizer
176,336 AstraZeneca
Text Mining
WordClouds (wordcloud)
N-grams (nltk)
LDA Topic Modelling (sklearn)
Sentiment Analysis
1. Polarity & subjectivity measures were assigned to the tweets (TextBlob)
2. Neutrality, positivity & negativity was assigned to the tweets (VADER)
EDA & Social Network Analysis
1. Processing of data for visualization (matplotlib)
2. Creation of networks with Gephi

Figure 1. Proposed methodology.

This work is presented as follows: Section 2 presents related work with focus on sentiment classification on social media posts, studies related to Twitter data and the COVID-19 disease. Section 3 describes the methods used to carry out this work, focusing on the main tasks carried out such as data collection, data preparation, sentiment analysis, exploratory data analysis, and social network analysis. In Section 4, all the results obtained using modeling and analysis of data are presented, while in Section 5 the importance of previously stated results are explained and evaluated. Finally, Section 6 is a brief conclusion and summary of all the findings made.

2. BACKGROUND

Social networks have enabled people to express themselves and show their interest or disinterest over a wide range of topics and situations. such as specific brands, campaigns, or products. For instance, in the case of companies that involve sales, the nature of customers' comments or reviews that are published within social networks greatly affects how current and potential customers evaluate the company's products and decide whether to stay on the company or not. Furthermore, by making use of social networks' data, companies may identify these opinions and make decisions based on them to improve in target areas. Such approaches have been made between apparel brands [8], in the automotive industry [9], and in food chains [10], among others.

Data from social networks has also had several applications in sociological and psychological studies. Some of these applications include the application of sentiment analysis to study social and cultural aspects ranging from attitudes towards women [11], suicide [12], and literary studies [13], among others.

The current COVID-19 pandemic has also been the subject of studies that involve people's opinions and their interpretation. For instance, a study by Kumar et al. [14] focuses on analyzing public opinion about online learning during confinement due to the COVID-19 pandemic. Another article published by Toeppe et al. [15], presents an approach to evaluate the different emotions reflected among affective publics towards the current pandemic. For this study, data from Twitter was employed for the sentiment analysis. Similarly, Twitter data has been used to evaluate sentiments about the vaccines that have been developed against SARS-CoV-2. For example, in one of the studies [16], Twitter data was employed to evaluate the opinion of the Indonesian population about their vaccination scheme using tweets from January 2021. In another study [17], millions of tweets in English from February 2020 to December 2020 were analyzed to identify the emotions of the population on the upcoming vaccines for COVID-19.

While previous studies effectively use Twitter data for providing an insight into the response towards the different COVID-19 vaccines, vaccine-related tweets in Spanish have not been yet analyzed and the sentiments of the Mexican population towards the vaccines that are available within the country have not been explored either. For this reason, we present this study which focuses on the previous topic within the Mexican population.

3. METHODS AND DATA

3.1. Data Collection

The data used for this study was obtained using the Twitter Developer API. Tweets collected were filtered using the coordinates and radius of Mexico to obtain only the tweets from Mexican users. The vaccines selected for analysis were those acquired by the Mexican government to initiate the vaccination program in Mexico, which are: Pfizer-BioNTech, Sputnik, OxfordAstraZeneca, CanSino, and Sinovac. An individual dataset was created for each of the vaccines, initiating the collection of tweets from March 07, 2021, until April 21, 2021.

A total of 176,336, 135,294, 42,819, 24,271, and 62,532 tweets were gathered for the OxfordAstraZeneca, Pfizer-BioNTech, Sinovac, CanSino, and Sputnik vaccines, respectively. These tweets were then loaded into a Python Pandas data frame for further analysis.

3.2. Data Preparation

Once loaded into data frames, the tweets were cleaned by removing the duplicate tweets occasioned by re-tweets (RT), the hashtags, users, and unnecessary spaces in tweets. Afterwards, the tweets were translated to English to proceed with the Sentiment Analysis. This was carried out using the Googletrans Python library. After processing the data, the remaining number of tweets were of 31,905, 26,641, 7,511, 4,204, and 10,928 for Oxford-AstraZeneca, Pfizer-BioNTech, Sinovac, CanSino, and Sputnik vaccines, respectively.

3.3. Sentiment Analysis

After the datasets were processed, subjectivity and polarity measures were assigned to the tweets. This was carried out using the TextBlob Python library, using the Sentiment property.

Neutrality, positivity, and negativity measures were assigned to the tweets using the Valence Aware Dictionary for sEntiment Reasoning (VADER) [18] tool from the Natural Language Toolkit (NLTK) package in Python.

Valence Aware Dictionary for sEntiment Reasoning (VADER) is a sentiment lexicon-based sentiment analyst constructed to work with text/sentiment expressed in social media microblogs. Its lexicon includes common features to sentiment expression in English such as emoticons (":)" referring to a smiley face, or ":(" referring to a sad face), acronyms ("LOL" and "WTF"), and slang ("meh", "nah"). Therefore, VADER outperforms other lexicon-based approaches in social media sentiment analysis. For instance, the approaches based on polarity lexicons (LIWC, GI, ...) are not able to capture the sentiment from emotions or acronyms since they are only capable of generating binary polarity. Although this polarity problem is solved by using valence-based sentiment approaches such as ANEW, this approach also fails to cover for most common lexical features in social media. Also, since VADER is not a machine learning model, problems such as the requirement of large datasets, high computational cost derived from training/classification time, and having features that are not easily interpretable due to processing in black boxes, are not present using this method.

Even though it may seem that approaches such as Linguistic Inquiry and Word Count (LIWC) [26], Affective Norms for English Words (ANEW) [27] or General Inquirer (GI) [28] are not as efficient for sentiment analysis, VADER's base sentiment lexicon was derived from those lexical features given they are already rated. The newly introduced lexicon was validated using a Wisdom-ofthe-crowd approach and then cleaned through a series of evaluations and validations to maintain the quality. Moreover, some generalizable heuristics were also created to incorporate wordorder sensitive relationships between terms. These heuristics include punctuation (!) to increase the magnitude of the sentiment intensity, capitalization to emphasize a sentiment-relevant word, degree modifiers (such as "extremely", "super"), contrastive conjunction "but" pointing a shift in sentiment polarity and examining the trigram preceding a sentiment-laden lexical feature. This allowed reaching the gold-standard list of lexical features used in VADER, which makes it the most convenient choice to classify the sentiment of tweets in the present dataset.

The classified tweets were then separated into different arrays depending on their sentiment and a word cloud was created using the WordCloud generator from Python to visualize words predominating in tweets belonging to each of the sentiments.

3.4. Exploratory Data Analysis

For the exploratory data analysis, the data of each of the vaccine datasets was manipulated using the Pandas library from Python. The time series for the number of tweets per day was developed using the time date Python module and the visualization was created using the matplotlib library. Further manipulation of data was carried out to prepare data for building the tree map of the percentage of tweets by state. The main results of these maps were coupled with the sentiment data to analyze sentiments on the days and states with the most tweets.

3.5. Social Network Analysis

To identify the users with the most influence in each of the vaccine datasets, the data from the user replies were used to build a network. The features of 'User ID' and 'In Reply to User' were used to build a table with the list of edges, with each edge referring to a connection of a user that replied to another user. Each of these users was referred to with their Twitter IDs and was set as the nodes of the graph. Since for this analysis we were interested in identifying the users with the most influence causing negative replies, the tweets that were selected for the graph were only those that had a negative classification. A single graph was built for every vaccine dataset using the Gephi software, filtering the nodes by their in-degree and showing the labels only for the nodes with in-degree ≥ 5.

4. RESULTS

4.1. Sentiment Analysis and Text Mining

Every tweet in each vaccine's dataset was assigned one of three different sentiments: neutral, positive, and negative. The number of tweets belonging to each class was normalized and is shown in Figure 2, where a general tendency towards neutrality can be observed. Figure 2 shows that Sinovac (54%) is the vaccine with the highest percentage of neutral tweets followed by Sputnik (48%), Pfizer (42%), CanSino (39%), and AstraZeneca (35%). Moreover, CanSino (39%) has a higher prevalence of positive tweets as opposed to Sinovac (28%), which had the lowest percentage. Sputnik and AstraZeneca presented the same percentage (32%), which was slightly lower than Pfizer (37%). Lastly, AstraZeneca (32%) had the highest percentage of negative tweets, followed by Pfizer (20%), CanSino (20%), Sputnik (19%), and Sinovac (17%). These percentages indicate an overview of people's reception towards a specific vaccine, but the context of why those sentiments arose is also a major topic to inquire about. Furthermore, it is important to mention that, although the following experiments were carried out using all the vaccine datasets, only the results for the AstraZeneca vaccine are shown in this paper. The summarized results for the remaining 4 vaccines can be visualized in Table 1 and Table 2.

The visualization of keywords on the created word clouds, along with existing evidence, allow the extraction of context that aid in the interpretation of the results from the sentiment analysis. For instance, in tweets related to positive sentiments found in the AstraZeneca dataset (Fig. 3), topics where people posted their absence of symptoms after they received their dosages could be found. Also, the millions of vaccines that would be shared by the USA with Mexico, with an initial batch of 1.5 million according to the Mexican Secretary of Foreign Affairs Marcelo Ebrard [19], was shown in the word cloud as well.

On the other hand, negatively classified tweets (Fig. 4) discussed the most commented side effect, which was the thrombosis that started appearing in several cases in European countries and of which news was spread quickly. Additionally, people appeared to be complaining about the government's decision not to suspend its use in Mexico. Most of these topics, including "blood clots", "side effects", "European countries", and "link between AstraZeneca", can also be observed in the 2-gram (Fig. 5) and 3-gram.

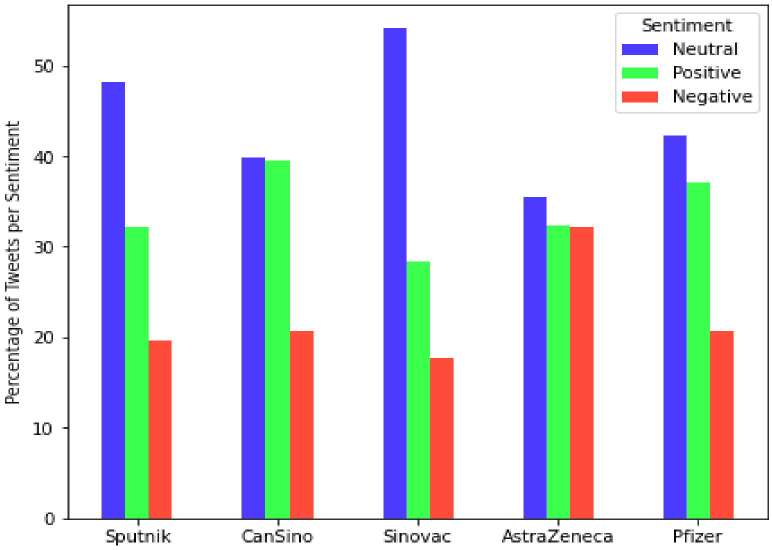

Figure 2. Obtained percentages of tweets classified by sentiment across five vaccines in Mexico.

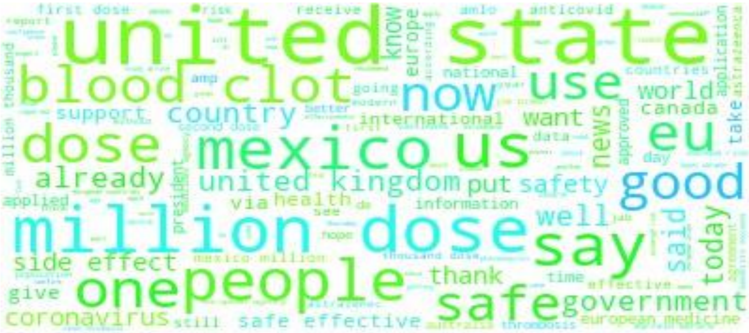

Figure 3. Word cloud based on positive AstraZeneca tweets.

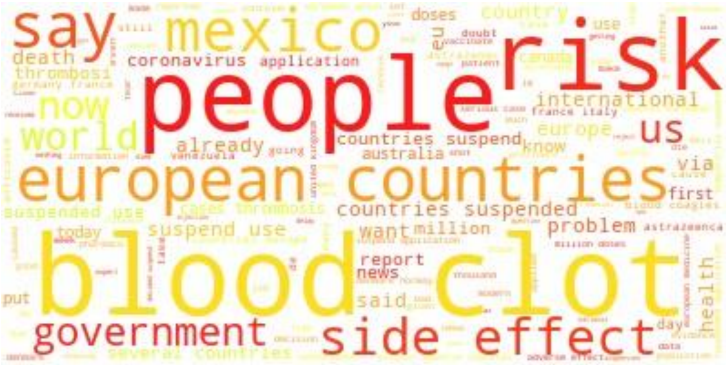

Figure 4. Word cloud based on negative AstraZeneca tweets.

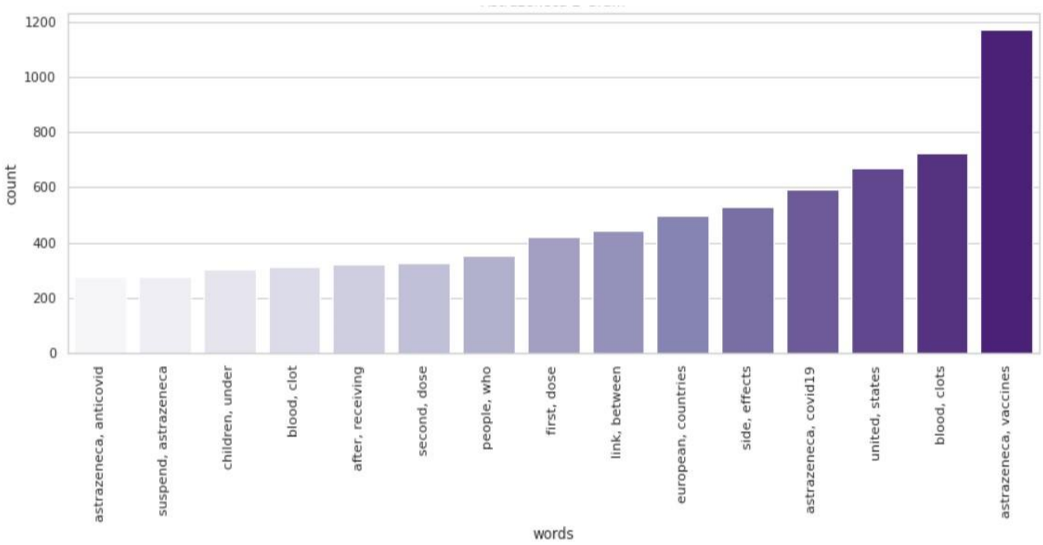

Figure 5. 2-gram on AstraZeneca vaccine dataset.

Similarly, the most important words appear in CanSino's positive word cloud, among which topics regarding its key feature of requiring only a single dose were found. Another topic involved people commenting about how their bodies reacted after receiving their unique dose. Being the most important news at that time was the packaging that would be carried out in Mexico at Queretaro's DrugMex Pharmaceutical Plant using CanSino's active substance [20]. Since this meant the vaccination progress would speed up, positivity in the tweets including this topic may be explained. On the other hand, tiredness was among one of the main symptoms people had days after their dose in connection to the negative sentiment tweets. In addition, negatively classified sentiments were linked to the vaccine's effectiveness, which was a very controversial issue when CanSino was first approved by the Mexican Federal Commission for the Protection Against Sanitary Risks (COFEPRIS). Likewise, those topics are detected in 2 and 3-Grams as "active substance", "single dose", "drugmex plant", "first three lots", and "over 60 years", where the latter refers to the decrease in the effectiveness in people over 60 years (Table 1).

In the case of the Russian vaccine Sputnik, negative sentiments also seemed to include several news. Among the most important news was the distribution of vaccines in the state of Campeche from apparently authorized retailers, in addition to the fact that, in the same state, vaccines were confiscated heading illegally to Honduras [21]. People also talked about their side effects after receiving the drug, but the previously mentioned news appeared to have more impact on the discussions. Similar to the previously discussed vaccines, the positively classified tweets from Sputnik, also showed people speaking up, either posting using their own Twitter profile or replying to other accounts, on the lack of symptoms or thanking governmental institutions for receiving their dose (Table 1).

Lastly, Sinovac's negative tweets referred to the users that were concerned over the vaccine's safety due to the death of an elderly person 40 minutes after he received his dose in the state of Guerrero. In other tweets, also about older adults, users were asking why Sinovac was still being applied to people over 60 years if only a few adults signed up to clinical trials, according to Europa Press news agency [22]. Besides this, another complaint was about the waste of vaccines due to the poor conditions in which they arrived in Mexico along with the bad temperature management that led to other vaccines being discarded as well. On the positive sentiment side, people talked about their experience days after receiving the vaccine and the millions of doses that were coming to Mexico [23] (Table 1).

Table 1. Sentiment analysis and text mining summarized results on CanSino, Pfizer, Sinovac, and Sputnik vaccine datasets.

Vaccine	WordCloud (Positive)	WordCloud (Negative)	2-Gram	3-Gram
CanSino	single, dose, teacher, china, packaged, queretaro, ebrard, thank, health, advantage.	tired, teacher, emergency, cofepris, authorize, effectiveness.	single, dose; tired, vaccines; older, adults; first, batch; marcelo, ebrard; drugmex, plant; emergency, use; active, substance.	Over, 60, years; authorizes, emergency, use; first, three, lots; Coahuila, Chiapas, Nayarit.
Pfizer	us, first, vaccine, second, dose, well, thank, time, better, already, health, want, effectiveness.	shot, got, already, days, problems, doses, death, case, risk, side, effect.	Second, dose; first, dose; older, adults; third, dose; 60, years; astra, Zeneca; side, effects.	Over, 60, years; my, first, dose; south, african, variant; Mexico, new, lot; children, under, 12; young, people, between.
Sinovac	well, china, effective, thank, today, older, without, apply, reached.	effectiveness, poor, condition, tired, problem, hidalgo, temperature, adult, death, serious.	Second, dose; older, adults; first, dose; 60, years; after, receiving; pharmaceutics, vaccines, Pfizer; arrived, at.	Adults, over, 60; Mexico, has, received; minutes, after, receiving; has, an, effectiveness; 15, minutes, later.
Sputnik	dose, well, president, health, already, thank, according, effective, come, hope, case.	false, Campeche, report, Honduras, doses, via, case, problem, government.	Second, dose; first, dose; false, vaccines; not, know; would, have; private, aircraft; astra, Zeneca; new, lot; alberto, Fernandez; false, doses; direct, investment.	Direct, investment, fund; v, vaccines, confiscated; private, aircraft, at; has, an, effectiveness; san, pedro, sula; Campeche, international, airport; European, medicines, agency; 5000, false, doses.

4.2. Exploratory Data Analysis

4.2.1. Number of Tweets per Day Time Series

In the Exploratory Data Analysis (EDA), the manipulation of data to obtain a time series of the count of tweets per day for each of the vaccines allowed us to identify trends within data. The number of tweets about the vaccines seemed to decrease on the weekend, resulting in a weekly trend of tweets having its peak in the middle of the week. However, a specific peak that corresponds to the day with the highest number of tweets per dataset could also be identified.

Most of the vaccines had the highest number of tweets in the middle of March (Table 2). For instance, the AstraZeneca dataset had the highest number of tweets on March 16 (Fig. 6) and the CanSino dataset on March 23. Also, the Sinovac and the Sputnik vaccine datasets both had their peak in the number of tweets on March 18, and, while the Pfizer dataset had a high number of tweets around these days, its peak was found on April 9.

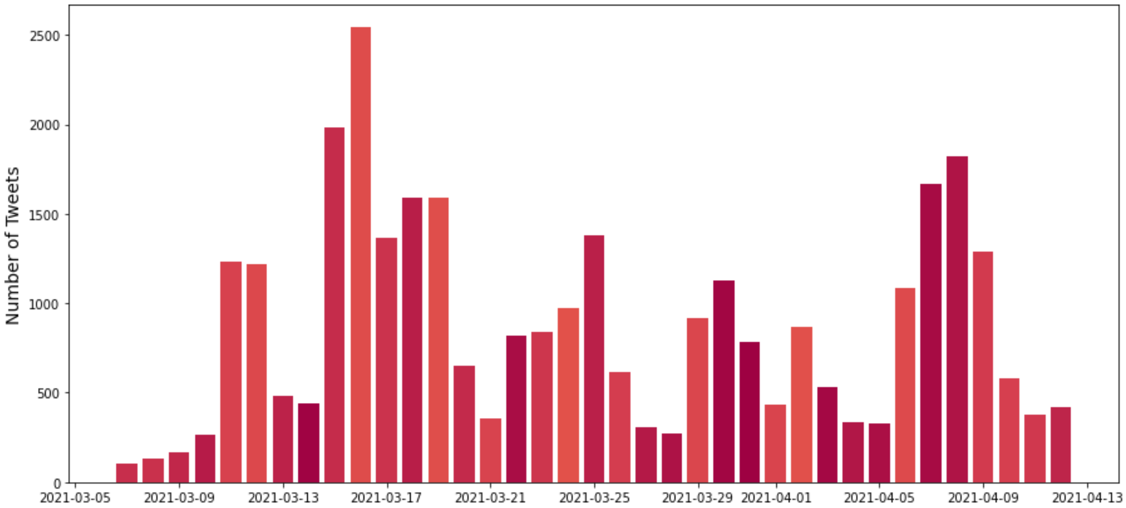

Figure 6. Number of AstraZeneca vaccine dataset tweets by day.

By tracing back on the events of the peak dates, it is possible to identify which type of events or news triggered a higher response on the Mexican population. Furthermore, by identifying the sentiments among the tweets on these peak dates, the nature of the response can also be elucidated. For instance, the AstraZeneca vaccine had a predominance of negativity on the day with the most tweets (Fig. 7). This was the same case for the CanSino vaccine, although the difference in the percentage of negative tweets was higher in the CanSino dataset (45%) than in the AstraZeneca dataset (38.5%). The remaining three vaccines had predominance for neutrality, resulting in percentages of 70.71, 56.56, and 38.88, for Sinovac, Sputnik, and Pfizer, respectively. In these cases, the percentage of neutral tweets was significantly higher in the Sinovac and the Sputnik vaccine datasets than in the Pfizer dataset, and, also, the second most predominant sentiment was negative in Sputnik and Sinovac datasets as opposed to the latter, in which there was a higher percentage of positive tweets (Table 2). This provides evidence that the immediate response to events about news for each of the vaccines had a positive nature for the Pfizer vaccine and a negative one for the rest of the vaccines, specially CanSino and AstraZeneca.

4.2.2. Percentage of Tweets by State

Data was further manipulated to view how the percentage of tweets obtained for each of the vaccine datasets was divided into each of the Mexican states. The results were quite similar among the five datasets (Table 2). The states of Puebla, Querétaro, and San Luis Potosí had the highest percentage of tweets in the AstraZeneca (Fig. 8) and the Sputnik datasets. Similarly, CanSino and Sinovac also had among their top 3 the states of Puebla and Querétaro, differing only by the states of Guanajuato and Oaxaca, respectively. Finally, the Pfizer vaccine dataset also had Puebla and Oaxaca among its top 3 states, along with Veracruz. These results indicate that the states that seemed to have the most activity on Twitter regarding the vaccines were Puebla and Querétaro.

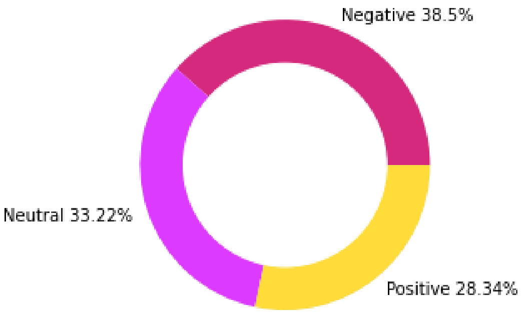

Figure 7. Percentage of tweets per sentiment classification on AstraZeneca vaccine dataset on March 16.

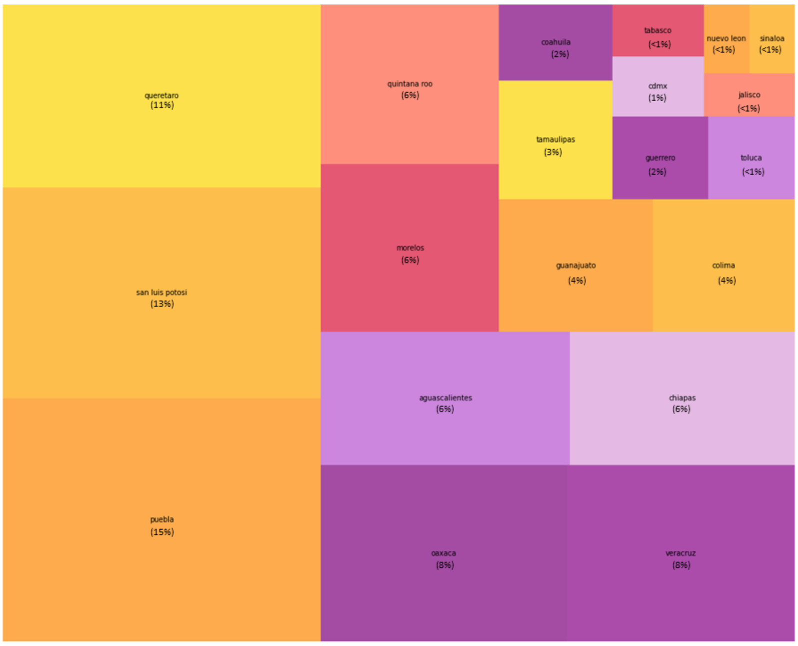

Figure 8. Tree map of percentages of tweets per state on AstraZeneca vaccine dataset.

Furthermore, the sentiments obtained were also coupled with these results (Table 2). For instance, the predominant sentiment on the top 3 states with a higher percentage of tweets was negative for the AstraZeneca vaccine dataset (Fig. 9). This was the same case for the state of Guanajuato on the CanSino dataset, which was also among the top three states with the highest percentage of tweets for this vaccine. However, the rest of the vaccine datasets and their respective top 3 states had a high predominance for neutral sentiments; and for all the vaccines except AstraZeneca and CanSino, the second most predominant sentiment was positive. These results may indicate that the overall response of the most active states regarding the vaccines was negative for AstraZeneca and CanSino and positive for Sputnik, Pfizer, and Sinovac.

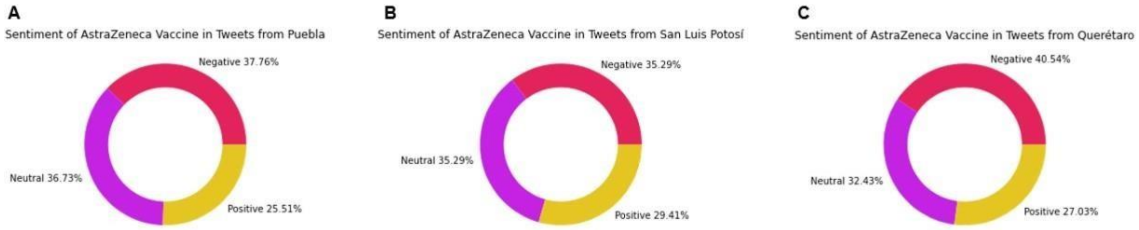

Figure 9. Percentage of tweets per sentiment classification on top three states with the highest tweet count on AstraZeneca vaccine dataset.

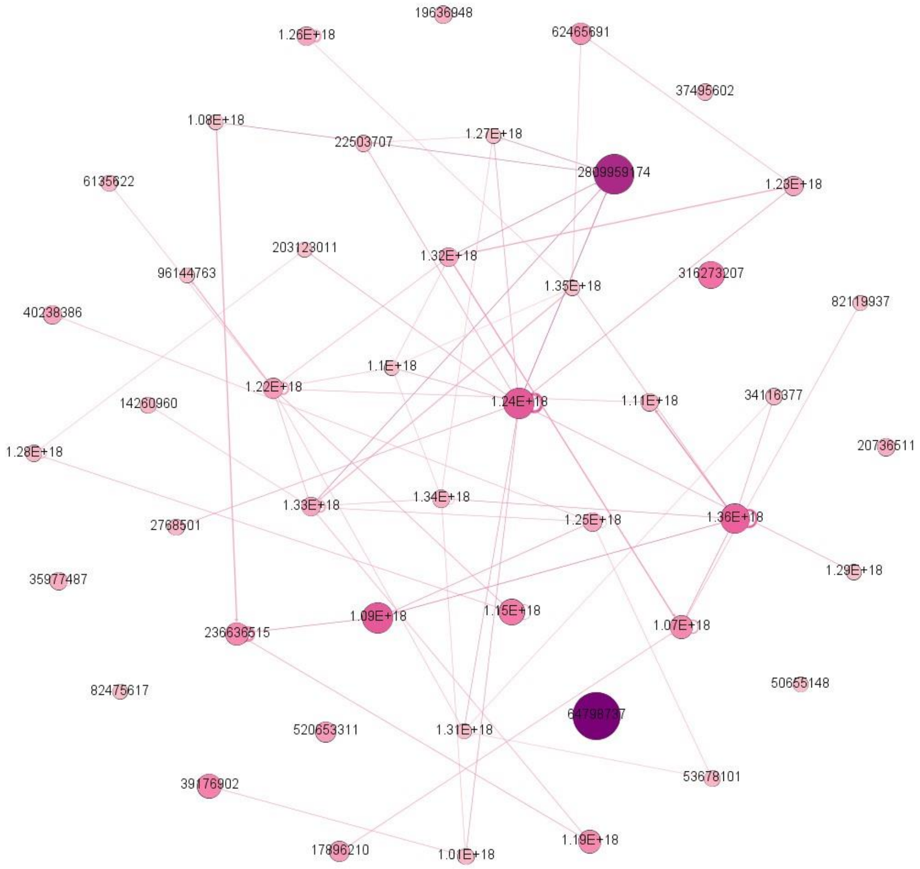

Figure 10. Social Network Analysis of user replies in AstraZeneca vaccine dataset. Graph is filtered by in-degree ≥ 5, in which labelled nodes correspond to this equality. Numbers in nodes represent the ID of the Twitter users: 64798737 is Marcelo Ebrard, 316273207 is Dr. Alejandro Macias, and 236636515 is Joaquin López-Dóriga.

4.3. Social Network Analysis

The Social Network Analysis was based on the user replies from each of the vaccine datasets. A network was generated per vaccine and was conformed of tweets that were classified as negative. Filtering the network allowed the identification of the users that had the most negative replies about the vaccines. For example, the Mexican Secretary of Foreign Affairs, Marcelo Ebrard, was one of the users with the most replies in all the vaccine datasets. Similarly, Dr. Alejandro Macías, a researcher from the University of Guanajuato, appeared as one of the nodes with the most replies in the AstraZeneca (Fig. 10), Sinovac, Sputnik, and CanSino vaccines. A famous Mexican journalist called Joaquín López-Dóriga was also highly present in the negative replies from the CanSino and AstraZeneca vaccine datasets.

Other figures such as researchers and politicians were also quite present in the vaccine datasets. For instance, Lilly Téllez, a Mexican journalist and politician had several replies in the Sputnik dataset. Also, the Quintana Roo Governor Carlos Joaquín appeared as a large node in the Pfizer dataset, while Dr. Alma Maldonado, a researcher, and professor from the University of Mexico was highly present in the CanSino dataset. The results from these networks indicate that politicians and researchers are highly involved in negative responses regarding each of the available vaccines in Mexico. Further investigations on the nature of the tweets that were highly replied to may provide an insight into the type of allegations that have the greatest impact on the target users.

Table 2. EDA and Social Network Analysis summarized results for 5 vaccine datasets.

	Sinovac	**Pfizer**	**CanSino**	**Sputnik**	**AstraZeneca**
Mean (tweets/day)	166	605	93	237	856
Max tweet count	478	1273	622	838	2507
Day with highest tweet count	March 18, 2021	April 9, 2021	March 23, 2021	March 18, 2021	March 16, 2021
State with highest tweet count	Puebla	Puebla	Puebla	Queretaro	Puebla
Sentiment (highest tweet count day)	Neutral	Neutral	Negative	Neutral	Negative
Sentiment (highest tweet count state)	Neutral	Neutral	Neutral	Neutral	Negative

Main subjects in social network	Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher)	Marcelo Ebrard (Politician); Carlos Joaquín (Politician)	Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher); Dr. Alma Maldonado (Researcher); Joaquin Lopez-Doriga (Journalist)	Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher); Lilly Téllez (Politician/Journalist)	Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher); Joaquin LopezDoriga (Journalist)

5. DISCUSSION

The analysis carried out in this research demonstrates an overview of people's reception towards a specific vaccine. The classes assigned to each tweet in every dataset allowed the visualization and analysis of the different types of reactions that were encountered for each of the vaccines, and the reason of those sentiments could be modelled using word clouds that enabled the identification of keywords that could provide a clue on the topic that was being discussed. For instance, side effects appear frequently among negative sentiment keywords, as well as complaints against the government, or reactions to the most shocking news affecting the vaccines, not only in Mexico but also in other parts of the world. On the other hand, positive tweets involved appreciation to the government for supplying the vaccines and spoke positively about the safety and effectiveness of the different vaccines.

To provide more context to the keywords found on the word clouds, 2-grams, and 3-grams for each of the vaccine datasets were also created. These n-grams allowed us to visualize the most important phrases from which a context can be extracted without the need for modelling using a Topic Model such as non-negative matrix factorization (NMF) or Latent Dirichlet Allocation (LDA). However, LDA was performed as a validation method for the relationship found between the topics on the word clouds and the n-grams. Given their similarity, LDA results are not shown in this paper. Furthermore, since the results of the n-grams were not divided into positive and negative tweets as in the word clouds, the small phrases that appeared depended on how positively or negatively people talked about the specific vaccine. For instance, since AstraZeneca had more negative tweets, the most recurrent n-grams were related to the blood clots side effect, while the case of CanSino, which was the vaccine with the most positive tweets had good comments regarding its single-dose feature. The identification of these sentiments and topics is relevant as social networks have a tremendous impact on mass behavior and tendencies. Consequently, although positive comments could have the effect of promoting vaccination, negative comments could also affect the decision of the population not to apply their vaccine.

A brief EDA was carried out to have a general visualization of how the tweets from each vaccine dataset were distributed. First, with a segmentation of the count of tweets per day, we were able to obtain the days with the most tweets of any sentiment for each of the vaccine datasets. By tracking back on the news or important announcements, the tweets from our database, and the results from the previous analyses, we could identify several situations in which the users reacted and tweeted the most. For instance, in the case of the Sinovac vaccine, the day with the most tweets,was the day where Mexico received a batch of a million doses [23], for which people had more positive than negative reactions. This was also the case for the Pfizer vaccine, where the main announcements involved the arrival of a new batch of vaccines, and the affirmation of the President that with this new arrival, the doses for all Mexican older adults would be sufficed

[24]; the tweets on this day were also mostly positive. Moreover, in the case of CanSino, announcements that were made on the day with the most tweets were related to the liberation and immediate application of the vaccines in rural regions and the acquisition of more CanSino doses that were packaged in the state of Querétaro [25]. As opposed to Sinovac and Pfizer, these announcements had a higher percentage of negative responses. Similarly, on the day with the most tweets on the AstraZeneca vaccine, the main announcement was the negotiation of Mexico with the US to obtain these vaccines [19], which also had a mostly negative response. Finally, the day with the most tweets in the Sputnik dataset matched the day of the announcement of the discovery of fake vaccines in the state of Campeche [21]. Although most of these tweets had a neutral tone, the negative responses predominated over positive ones. This information leads to the conclusion that overall, depending on the previous "reputation" of the vaccine, announcements involving the arrival of more doses would define the polarity of users' reactions towards it.

Also, the datasets were used to explore the states that tweeted the most about each of the vaccines. Overall, it was found that the number of tweets was concentrated in the southeast and north-central regions of Mexico. While most results indicated neutrality and positivity towards the vaccines, there was a notorious exception in the case of the AstraZeneca and CanSino vaccines, whose tweets from their top three states with the highest number of tweets had a negative-neutral tendency. While these results provide evidence on the regions of the country that are the most active in Twitter regarding the vaccines, it would be necessary to further study the nature of the tweets by each of the states to define the impact that these states are having on the rest of the country. Furthermore, it would be necessary to validate whether the datasets are balanced in terms of the number of tweets that were recovered by state using the API.

The results from the Social Network Analysis let us identify those politicians, journalists, and researchers that were the most influential figures on the vaccine datasets. For instance, the Mexican Secretary of Foreign Affairs, Marcelo Ebrard, was the node with the highest in-degree in almost all the datasets. This was expected as the Secretary gives several announcements on the negotiations and transportations of the different vaccines. A similar effect was visualized in the case of the famous Mexican journalist Joaquin López-Dóriga. However, politicians Lilly Tellez and Carlos Joaquín, as well as researchers Alejandro Macias and Alma Maldonado, could represent specific targets within the datasets whose comments trigger a response on the community. Since the networks were built based on the negatively classified tweets, we can assume that these people triggered a negative response, however, deeper analysis on each of the subjects to define what is the exact nature of their impact and which characteristics they present to create this impact in the community. It is important to mention that, although targets such as these could be extracted from the networks, noisy users were also found. For instance, in the case of the Sputnik vaccine, one of the nodes with the highest in-degree was that of an influencer named "Sputnik". These types of cases introduce noise to the analysis and should therefore be considered in future work.

Overall, the sentiment analysis coupled with the text mining, the EDA, and the social network, allowed us to have a deep overview of each of the datasets created for each of the vaccines available in Mexico. This approach resulted useful to gain insight into the reactions and behaviors of the Mexican population towards the vaccines and to open the landscape to future investigations where misinformation sources can be predicted, and to further implement measures that may slow down the spread of fake news or allegations that may be harmful to the community. Actions that can be taken to further improve the current method include refining the geolocation of the queries from Twitter, curating users that may include noise into the datasets, including a wider variety of sentiments for classification, and creating social networks that demonstrate the polarity of the community. Further work on vaccine follow-ups is also encouraged.

6. CONCLUSION

The method employed in this work for evaluating the sentiments of Mexicans towards the currently available COVID-19 vaccines was effective in gaining an insight into the behavior and opinions of the Mexican population. Using the VADER toolkit allowed the classification of tweets into three different sentiments, providing a format with which the data could be further analyzed. This method combined with text mining, exploratory data analysis, and social network analysis allowed the identification of key words, dates, geographical regions, and people that were highly represented by the data. While improvements such as expanding the number of sentiments from the tweets and refining translation can be made, the current analysis provides a framework that contributes to the understanding of population behavior and to the identification and avoidance of possible misinformation.

REFERENCES

[1] Daniel Allington, Bobby Duffy, Simon Wessely, Nayana Dhavan, and James Rubin. Health-protective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency. Psychological Medicine, page 1–7, 2020.

[2] World Health Organization. Ten threats to global health in 2019. https://www.who.int/newsroom/spotlight/ten-threats-to-globalhealth-in-2019, Mar 2019.

[3] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. PloS one, 5:e14118, 11 2010.

[4] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Viral misinformation: The role of homophily and polarization. In Proceedings of the 24th International Conference on World Wide Web, WWW ’15 Companion, page 355–356, New York, NY, USA, 2015. Association for Computing Machinery.

[5] J. Leung, J. Y. C. Chung, C. Tisdale, V. Chiu, C. C. W. Lim, and G. Chan. Anxiety and panic buying behaviour during covid-19 pandemic-a qualitative analysis of toilet paper hoarding contents on twitter. International Journal of Environmental Research and Public Health, 18:1–16, 2021.

[6] J. Talbot, V. Charron, and A.T. Konkle. Feeling the void: Lack of support for isolation and sleep difficulties in pregnant women during the covid19 pandemic revealed by twitter data analysis. International Journal of Environmental Research and Public Health, 18:1–12, 2021.

[7] A. Jabeen, S. Afzal, M. Maqsood, I. Mehmood, S. Yasmin, M.T. Niaz, and Y. Nam. Anlstm based forecasting for major stock sectors using covid sentiment. Computers, Materials and Continua, 1, 2021.

[8] Abdur Rasool, Ran Tao, AlimarjanKamyab, and Tayyab Naveed. Twitter sentiment analysis: A case study for apparel brands. volume 1176, page 022015, 03 2019.

[9] Sarah Shukri, Rawan Yaghi, Ibrahim Aljarah, and Hamad Alsawalqah. Twitter sentiment analysis: A case study in the automotive industry. 11 2015.

[10] Wu He, ShenghuaZha, and Ling li. Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33:464–472, 06 2013.

[11] Muna Al-Razgan, Asma Alrowily, Rawan L. Al-Matham, Khulood M. Alghambdi, MahaShaabi, and Lama Alssum. Using diffusion of innovation theory and sentiment analysis to analyze attitudes toward driving adoption by saudi women. Technology in Society, 65, 05 2021.

[12] E. Rajesh Kumar and K.V.S.N. Rama Rao. Sentiment analysis using social and topic context for suicide prediction. International Journal of Advanced Computer Science and Applications(IJACSA), 12, 2021.

[13] Evgeny Kim and Roman Klinger. A survey on sentiment and emotion analysis for computational literary studies, 2018.

[14] Kaushal Kumar Bhagat, Sanjaya Mishra, Alakh Dixit, and Chun-Yen Chang. ”public opinions about online learning during covid-19: A sentiment analysis approach. Sustainability, MDPI, 13:1–12, 03 2021.

[15] Katharina Toeppe, Hui Shenghua Yan, and Samuel Kai Wah Chu. Repurposing sentiment analysis for social research scopes: An inquiry into emotion expression within affective publics on twitter during the covid-19 emergency. Diversity,Divergence,Dialogue, page 396–410, 02 2021.

[16] Pristiyono, MulkanRitonga, Muhammad Ali Al Ihsan, AgusAnjar, and Fauziah Hanum Rambe. Sentiment analysis of covid-19 vaccine in indonesia using naıve bayes algorithm. IOP Conference Series: Materials Science and Engineering, 1088, 2021.

[17] Sameh N. Saleh, Samuel A. McDonald, Mujeeb A. Basit, Sanat Kumar, Reuben J. Arasaratnam, Trish M. Perl, Christoph U. Lehmann, and Richard J. Medford. Public perception of covid-19 vaccines through analysis of twitter content and users, 04 2021.

[18] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015.

[19] ’exitosa’ la negociacion con eu para que mexico obtenga vacunas de astrazeneca: Lopez-gatell, Mar 2021.

[20] Marittza Navarro. Drugmex, la empresa en queretaro que envasara la vacuna china contra el covid19, 02 2021.

[21] ElıasCamjahi and Marıa R. Sahuquillo. Decomisadas en mexicomas de 5.000 dosis falsas de la vacuna rusa sputnik v.

[22] MSN Noticias. La omsaprobo el uso de emergencia de la vacuna de sinovac, 06 2021.

[23] Cesar Arellano Garcıa. Llega a mexico un millon de vacunas de sinovac, Mar 2021.

[24] Mexico cuenta con todas las dosis necesarias para terminar de vacunar a personas adultas mayores el 20 de abril: presidente, Apr 2021.

[25] Vacunas cansino se usaran inmediatamente en los 32 estados: Lopezgatell, Mar 2021.

[26] Pennebaker, James & Francis, Martha & Booth, Roger. Linguistic inquiry and word count (LIWC). 1999.

[27] Bradley, M. and P. Lang. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. 1999.

[28] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates.The General Inquirer: A Computer Approach to Content Analysis.MIT Press, 1966.

AUTHORS

Jessica Salinas is a Biotechnology Engineer from Monterrey, Mexico, currently pursuing a master’s degree in Computer Science at Tecnológico de Monterrey, Mexico. Her main work experience has been working in research laboratories and IT consulting at Accenture Technology, Mexico. Nevertheless, Jessica also has experience in teaching languages and in developing IT workshops for university students. Her current interests rely mainly in Bioinformatics applied to health and plant science, Data Science, Data Analytics, and Data Visualization.

Carlos Flores Munguia received the B.S. degree in Computer Systems engineering from Tecnologico Nacional de Mexico, Tamaulipas in 2019, graduating with honours thanks to achievements obtained during his university career. Carlos is currently earning a master’s degree in Computer Science at Tecnologico de Monterrey. He has served as a teacher of different workshops with the purpose of helping students to develop talents not explored by the institution. He has also participated and won in innovation competitions, for which he has represented his city at state level. His research interests include the success in the application of genetic algorithms in real world problems, high performance computing, and Few-Shot Learning as an alternative to data-intensive Deep Learning approaches.