Analysis in the Study of Emotion and Social Media Consumption in Puerto Rico

Adel del Valle Pérez
Department of Mathematical Sciences
University of Puerto Rico at Mayagüez
Mayagüez, Puerto Rico
adel.delvalle@upr.edu

Bryan Ortiz-Torres
Industrial Engineering Department
University of Puerto Rico at Mayagüez
Mayagüez, Puerto Rico
bryan.ortiz4@upr.edu

Keywords—data mining, social networks, clustering, sentiment analysis

I. Abstract

Internet has revolutionized the way that scientists in different disciplines collect data. Advances in technology and the tendency of using social media to express opinions and emotions in real time has provided an opportunity to study how different events influence these opinions. Additionally, literature suggests that social media consumption increments in emergency events. This report presents preliminary results in the study of social media consumption and its relationship with how positive or negatives were the reaction of Twitter users in Puerto Rico from November 2019 to March 2021, where it is known that the country confronted earthquakes and the Covid-19 pandemic. Also, this project makes use of unsupervised learning techniques to cluster users based on topic categories. Preliminary results show that there was an increment in social media consumption and that users tend to cluster in two different groups.

II. Introduction

The internet has opened a massive door for people to connect in ways never imagined. In the past, acquiring experimental data was a challenge for researchers who aimed to study social interaction due to the lack of technological mechanisms. From waiting weeks to receive a letter, people can now chat and interact instantly in the virtual hyper-connected reality. Alongside the advancements in computational power and distributed computing, scientists today are able to extract the millions of data produced per second in social media to perform analysis and implement Machine and Deep Learning algorithms.

Social media consumption has been increasing during recent times, especially during crisis and emergency situations [12]. Millions of people have shifted from traditional media channels such as newspapers to online communication channels. Thanks to the accessibility of publishing news, opinions, reactions to different topics and learn information, most of the social media platforms have played an important role spreading awareness of natural disasters, political situations and social movements around the world.

Most of the literature that has studied Online Social Media has been done with Social Network Analysis. Literature in this topic stands by a range from individual traits to biggest relational traits that are shared between different sections of the network. These collections aggregate into network motifs. As users are free to interact with each other, they tend to form subgroups with ranging edges users being more interconnected with one another versus others [1]. This causes a self-organized network - where nodes have the freedom to emerge and remove connections - that share a common "small world" structure [2].

This project aims to extract and collect social media data from Twitter API to perform an exploratory analysis from a given interval of time. The expectation is to group users from Puerto Rico based on different topics and model relationships with the intention of discovering insights and patterns associated with emotional states. Specifically, this project seeks to use sentiment analysis and unsupervised learning with the implementation of clustering algorithms.

Extracting data from this social network platform represents a computational challenge due to large amounts of data to process, accessibility to a representative sample of tweets and users, filtering data from a specific location and acquiring tweets from a given interval of time. Also, sentiment analysis might confront difficulties when analyzing words due to the absence of packages or libraries to process other languages like Spanish.

Analyzing data from users in Puerto Rico could also contribute to model how emotions change across the year where the people in the country confront different challenges like earthquakes at the end of 2019, the pandemic caused by Covid-19 and the elections that took place in November 2020. However, it is important to recognize that even if the amount of data collected is representative of the Twitter community in Puerto Rico, inference may not apply to the entire population in the country.

III. Literature Review

Social Networks, considered a primary source of human behavior, are represented as a series of nodes and edges where the nodes represent individuals, and the edges represent the relationship between different individuals [6][1]. Edges can be classified into directed or undirected depending on how people are able to interact. In addition, there are two essential concepts to understand the behavior of social networks: homophily and the principle of triadic closure. Homophily refers to the idea that nodes who are connected to one another are more likely to have similar properties, and is used in many applications like node classification. These properties may include not only common backgrounds from two or more individuals, but beliefs, hobbies and other interests. On the other hand, the triadic closure principal details that the likelihood of being connected now or in the future is higher if two individuals have a friend in common. In other words, this principle implies an inherent tendency of real-world networks to cluster and consequently, a correlation in the edge structure of the network. Therefore, Social Network Analysis could reveal important information, flow and patterns that could be used for decision making. Among the applications that can be used for Social Network Analysis is Sentiment Analysis.[3]

Sentiment analysis, also known as Opinion Mining or Emotion AI, is a well-known natural language processing methodology that classifies data as positive, negative or neutral with the intention to analyze subjective information by extracting and identifying affection states. This methodology is widely used in marketing research because it provides insights of reactions people have to different variables. Even with the advances of technology, this technique confronts many challenges when trying to define what is a positive or a negative emotion. [4] [5]

Sentiment Analysis literature work focuses on different approaches that have been used: a Machine Learning approach and a Lexicon-based approach. Figure 1 shows specifications of these approaches.

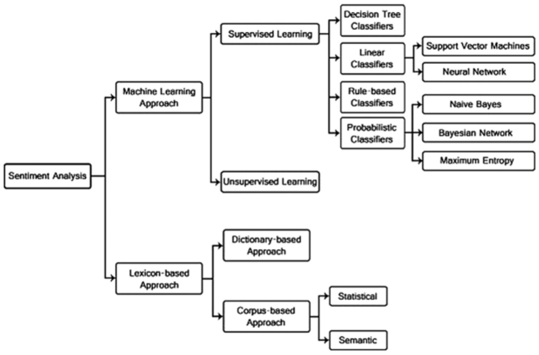


Fig. 1. Sentiment Analysis Approaches and Techniques. [5]

Data from Twitter platform has previously been used for different analyses because of the wide variety of subjects expressed by searching hashtags and words related to those topics, mostly using Social Network Analysis. This platform is affected by geographical and language diversity and has shown that people who discuss the same language tend to have similar clusters [6]. Additionally, it allows to track relationships in the cases where a node is a follower of another node. Therefore, this platform can be useful to understand public discussion, political and social movements, perceptive of consumer in different products, tourism, early warning of tsunamis, and many more.

Gupta et al. identified communities from users posting messages on Twitter during crisis events. In the study, three major crisis events of 2011 where considered: hurricane Irene, the riot in the United Kingdom and the earthquake in Virginia. These events provoked a vast number of posts in social media during and after the events. Results show that after identifying the top central users or people, to understand a community, there is a need to monitor and analyze only these top users rather than all the users in a community.

Jastania et al. [7] used Social Network Analysis to study the Arabic public discussion by extracting two million tweets from about 680,000 users and dividing the dataset into several types of networks: Retweet, Mention-Network, Co-Mention-Network and Hashtag-Network. In this way, this study used the metrics of graph centrality to reveal essential people in the discussion and was able to identify influencers.

In addition, in 2008, Diakopoulos and Shamma [8] used Sentiment Analysis to understand reaction of the audience during the U.S. presidential debate. For this, they extracted tweets from a given interval of time to study the tenor of the tweets during the debate. Tweets were rated in four different categories: *positive*, *negative*, *mixed* (contains both positive and negative components) and *other* (to catch statements unable to be classified). Also, they study if the tweets were favoring a specific candidate and the Pearson correlation between given topics of the debate and positive and negative responses.

Fornacciari et al. [9] present a combined approach between Social Network and Sentiment Analysis. For this, they collected three types of data from #SamSmith channel during the Grammy Awards in 2015 and the #Ukraine channel during the crisis of 2014: from users' profile, tweets from each user and follow relationships among users. Also, they used filters to eliminate useless tokens, special characters and symbols, performed orthographic corrections and used Multinomial Naïve Bayes, an algorithm which performs really well for text classification, to classify sentiments. In particular, they tried to associate sentiments to the nodes in the graphs of the social connections and to identify the hierarchy of the communities in the network.

IV. DATA

A sample of 226,000 tweets were collected from 403 different users. Table IV in *Appendix A* describe the data extracted using Twitter API.

V. METHODOLOGY

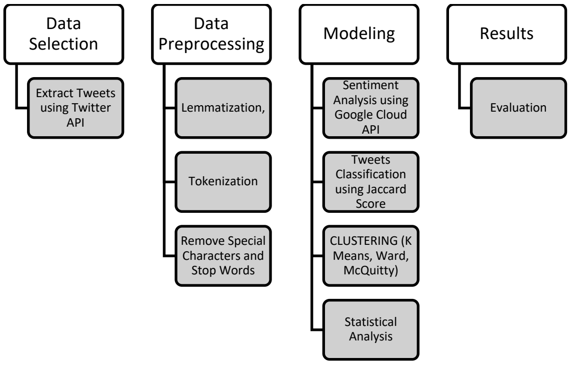

Data Selection
Extract Tweets using Twitter API
Data Preprocessing
Lemmatization,
Tokenization
Remove Special Characters and Stop Words
Modeling
Sentiment Analysis using Google Cloud API
Tweets Classification using Jaccard Score
CLUSTERING (K Means, Ward, McQuitty)
Statistical Analysis
Results
Evaluation

Fig 2. Methodology used in the analysis.

A. Data Selection

As Fig. 2. shows, the methodology followed started with extracting tweets. This extraction was entirely performed using Python open source and Twitter API. For this, a random sample of about 250,000 tweets from 403 users was collected.

This sample include all tweets published by the users from November 2019 to March 2021.

B. *Data Pre-processing*

The preprocessing procedure involved lemmatization, tokenization, removal of special characters and stop words and converting words into vectors. Lemmatization, in computational linguistics, is the process of determining the lemma or canonical form of a word based on its intended meaning and depends on correctly identifying the part of speech of a word in a sentence as well as the neighboring sentences [14]. Tokenization, on the other hand, refers to the process of demarcating and classifying sections of a string [15]. Removal of special characters consisted of deleting symbols that appeared on the tweets so that only letters are analyzed while stop words refers to remove common words that have no meaning or importance in the sentence. Both Spanish and English stop words were removed. Converting words into vectors (Word2Vec) is a Natural Language Processing technique that uses a neural network to learn word associations, detect synonymous words or suggest additional words for a partial sentence. This technique also represents each distinct word with a unique vector of numbers. Making it easier to use mathematical similarity measures [16].

C. *Modeling*

The modeling consisted of different stages where supervised methods were used to classify tweets while unsupervised methods were performed to group or cluster users. These stages were Sentiment Analysis, Classification of Tweets and Clustering.

VI. Results

A. *Data Pre-processing approaches*

Given that it was possible to extract all tweets from a user in an interval of time, there were no missing values in that stage of the pre-processing. However, in the sentiment analysis, after cleaning the data, some tweets were not scored due to the nature of the sentence (e.g., numbers, languages that were not Spanish or English). Therefore, they were removed.

B. *Sentiment Analysis Results*

After performing all the data preprocessing a sentiment analysis was performed on all the tweets analyzed using Google Cloud API. *Table V* in *Appendix A* shows a sample of the results from a specific user. Fig. 3 shows the frequency of tweets through time. The graph suggests that for the sample of users considered in the analysis, there was an exponential incrementation of tweets published from November 2019 to March 2021 with a brief reduction between November 2020 and January 2021. The graph also shows that at the beginning of November there was a period of time where the consumption of the social network in these users increased significantly compared to the pattern that was being observed.

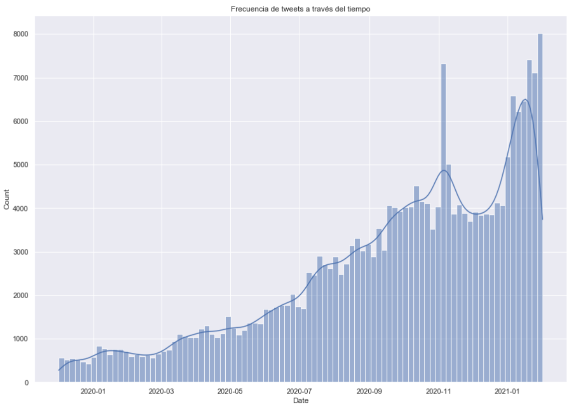


Fig 3. Frequency of tweets through time.

On the other hand, Fig. 4 presents the sum of the Sentiment Score by weeks. The graph behaves with same incremental pattern, yet, it can be seen that a maximum consumption of the social network is reached in the first weeks evaluated.

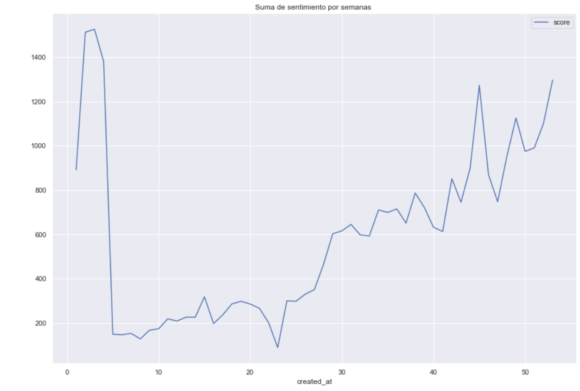


Fig. 4. Sum of Sentiments by Weeks.

Fig.5 shows an histogram of Sentiment Scores. The figure suggest that most of the tweets published by the users were scored above cero. Consequently, tendency shows that most of the tweets were positive. In addition, figure also illustrates that most tweets keep a score between 0.0 and 0.40.

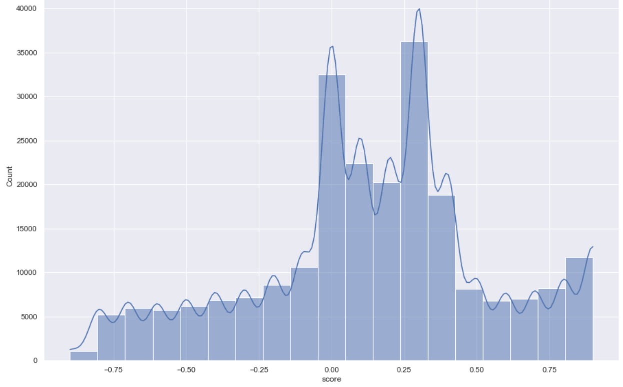
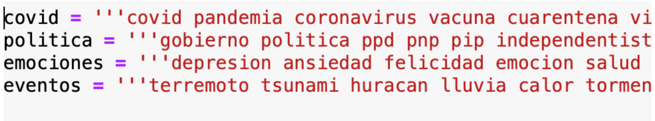


Fig. 5. Frequency of Sentiment.

C. *Classification of Tweets*

To classify the different tweets in categories of topics, a list of words related to each one of those topics was created. The Jaccard Score was performed on each tweet to classify the tweet with the most similar category. Fig. 6 shows a sample of the words used for each category.



```
covid = '''covid pandemia coronavirus vacuna cuarentena vi
politica = '''gobierno politica ppd pnp pip independentist
emociones = '''depresion ansiedad felicidad emocion salud
eventos = '''terremoto tsunami huracan lluvia calor tormen
```

Fig. 6. Sample of words used for the classification.

D. *Clustering*

Three different clustering methods were considered for this analysis: K Means Clustering, Ward Method and McQuitty Method. Dunn Index, Calinski-Harabasz Index, Davies-Bouldin Index, Silhouette Score and Duda Index performance measures were used to evaluate the performance of the clustering algorithms.

1) *Kmeans Clustering*

Fig. 7 to Fig. 9 provides a visual representation of the clusters generated using Kmeans. Using the metrics provided in Table #, the number of clusters suggested should be k=2 based on a voting scheme.

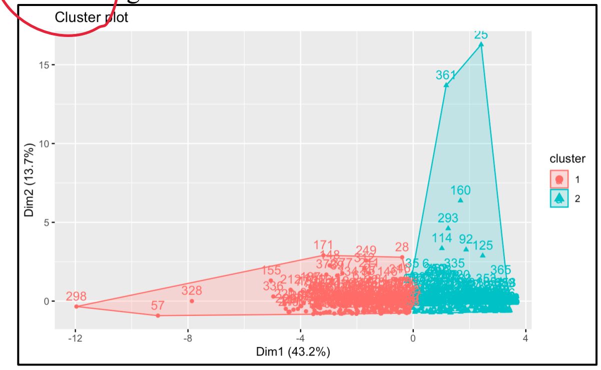


Fig. 7. Kmeans Clustering, k=2.

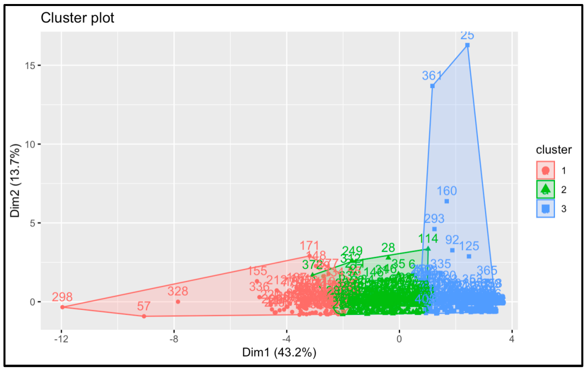


Fig. 8. Kmeans Clustering, k=3.

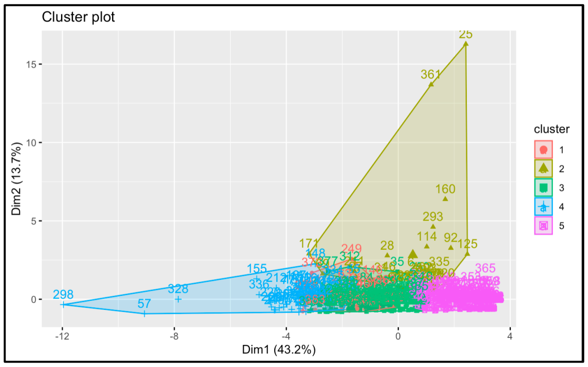


Fig. 9. Kmeans Clustering, k=5.

TABLE I. BEST K DETERMINED BY THE DIFFERENT METRICS IN KMEANS

Metrics	Best (K)
Dunn	16
CH	2
DB	12
Silhouette	2
Duda	2

2) *Ward Method*

Fig. 10 to Fig. 12 provides a visual representation of the clusters generated using Ward Method. Using the metrics provided in Table #, the number of clusters suggested should also be k=2 based on a voting scheme.

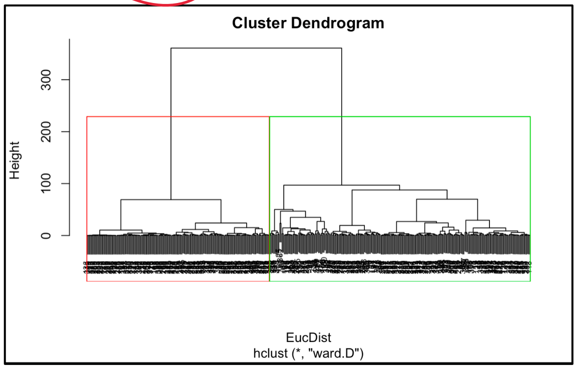


Fig. 10. Ward Method Clustering, k=2.

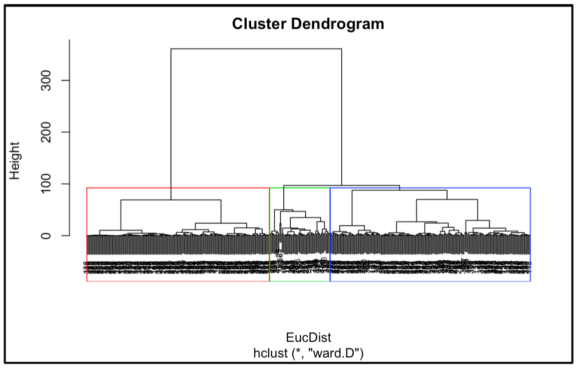


Fig. 11. Ward Method Clustering, k=3.

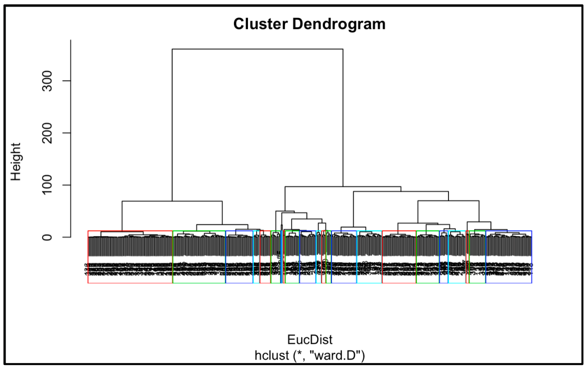


Fig. 12. Ward Method Clustering, k=23.

TABLE II. BEST K DETERMINED BY THE DIFFERENT METRICS IN WARD METHOD

Metrics	Best (K)
Dunn	23
CH	2
DB	25
Silhouette	3
Duda	2

3) *McQuitty Method*

Fig. 10 to Fig. 12 provides a visual representation of the clusters generated using the McQuitty Method. The metrics provided in Table #, suggest that the number of clusters should also be k=2.

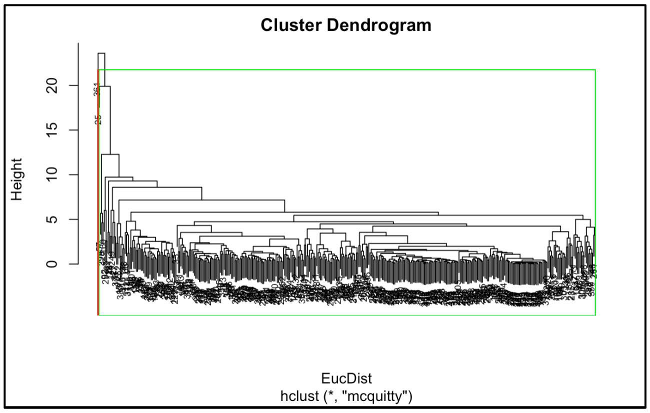


Fig. 13. McQuitty Method Clustering, k=2.

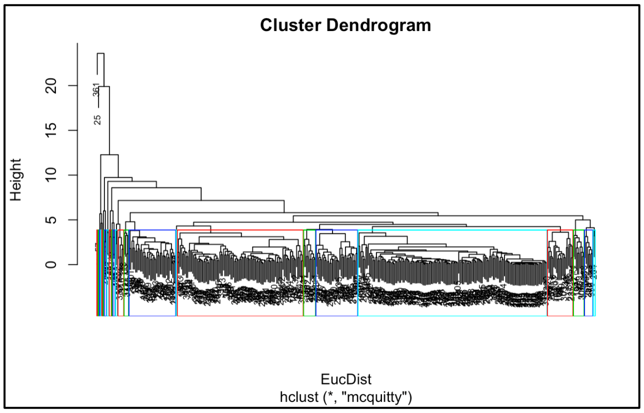


Fig. 14. McQuitty Method Clustering, k=23.

TABLE III. BEST K DETERMINED BY THE DIFFERENT METRICS IN MCQUITTY METHOD

Metrics	Best
Dunn	2
CH	24
DB	2
Silhouette	2
Duda	2

4) *Desirability Function & Best Model Selection*

To decide the best clustering technique, the following Desirability function was used. Results are shown in Table #. Given that the desirability function is based on the standard deviation of the scores, the option that minimizes desirability is chosen.

$$D_i = \sum \frac{\sigma_{\text{j}}}{\sqrt{k}} \tag{1}$$

Where D_i=Desirability of method I, σ = standard deviation of metric j and K = number of clusters selected as best.

TABLE IV. DESIRABILITY OF EACH METHOD CONSIDERED

Method	Desirability
Kmeans	16.65
Ward	10.05
McQuitty	5.13

VII. CONCLUSIONS

Performing an exploratory analysis on the sample of tweets provided significant information. As the results suggest, the Consumption of Social Networks increased in the established period of time. However, during that period, Puerto Rico was facing different points of emergency or crisis. Among them are the earthquakes that occurred between the period of November 2019 and February 2021, where many people were affected by their belongings, their homes and the public education system stopped the start of classes. On the other hand, it is found that in mid-March, the pandemic caused by Covid-19 in the world kept the entire country working and studying from their homes. The restrictive measures taken by the government were conservative and this could somehow influence the high consumption of social networks. An interesting detail is that despite the fact that in all the time considered the country was facing two emergency situations, the Sentiment Analysis scores tended to be more positive than negative.

In addition, the cluster analysis carried out suggested, in the three different methods, grouping the users into two different groups (k = 2). It is recommended to carry out inferential analyzes to identify if there is indeed a relationship between the variables analyzed, sentiment and the increasing trend of social media consumption.

The results of this study are preliminary. Nevertheless, it is recommended to improve the word vectors used to classify tweets in different categories, to collect more data by adding new users and opening a higher time interval, and to perform other types of modeling (e.g. Regression, Association Rules) to further study these relationships.

REFERENCES

[1] D. J. Watts, Small worlds: the dynamics of networks between order and randomness. Princeton, NJ: Princeton University Press, 2004.

[2] S. Milgram, “The small-world problem,” . Psychology Today, 1967.

[3] C. C. Aggarwal, “Social Network Analysis,” in Data mining: the textbook, Cham: Springer, 2015, pp. 618–661.

[4] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, “Sentiment analysis and classification based on textual reviews,” in 2013 International Conference on Information Communication and Embedded Systems (ICICES), Feb. 2013, pp. 271–276, doi: 10.1109/ICICES.2013.6508366.

[5] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, “Sentiment analysis: A review and comparative analysis of web services,” *Information Sciences*, vol. 311, pp. 18–38, 2015.

[6] F. Ahmad and G. Widén, “Language clustering and knowledge sharing in multilingual organizations: A social perspective on language,” *Journal of Information Science*, vol. 41, no. 4, pp. 430–443, 2015.

[7] Z. Jastania, M. Ahtisham, R. Ayaz, and K. Saeedi, “Using Social Network Analysis to Understand Public Discussions: The Case Study of #SaudiWomenCanDrive on Twitter,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020.

[8] N. A. Diakopoulos and D. A. Shamma, “Characterizing debate performance via aggregated twitter sentiment,” *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, 2010.

[9] Fornacciari, P., Mordonini, M., & Tomaiuolo, M, “Social Network and Sentiment Analysis on Twitter: Towards a Combined Approach”. *KDWeb,* 2015.

[10] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: a survey, Ain Shams Eng. J. (2014).

[11] W. Tan, M. B. Blake, I. Saleh, and S. Dustdar, ―Social-network-sourced big data analytics,‖ IEEE Internet Comput., vol. 17, no. 5, pp. 62–69, 2013.

[12] Mendoza et al. Twitter under crisis: can we trust what we rt? In Proceedings of the First Workshop on Social Media Analytics, SOMA ’10.

[13] A. Gupta, A. Joshi, and P. Kumaraguru, “Identifying and characterizing user communities on Twitter during crisis events,” Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media - DUBMMSM '12, 2012.

[14] T. Müller, R. Cotterell, A. Fraser, and H. Schütze, “Joint Lemmatization and Morphological Tagging with Lemming,” *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[15] J. J. Webster and C. Kit, “Tokenization as the initial phase in NLP,” *Proceedings of the 14th conference on Computational linguistics -* , 1992.

[16] Mikolov, Tomas & Corrado, G.s & Chen, Kai & Dean, Jeffrey. “Efficient Estimation of Word Representations in Vector Space”. 2013.

APPENDIX A. DATA DESCRIPTION, STATISTICS & SENTIMENT ANALYSIS SAMPLE

TABLE V. DATA DESCRIPTION & STATISTICS

Variables	**Type**	**Description**	**Mean**	**Minimum**	**Maximum**	**Total**	**Total**
created_at	*string*	Date when account was created	-	6/24/08	1/15/21	-	403
followers	*integer*	User's followers count	1876.0	1	207411	756045	403
following	*integer*	User's following count	710.6	1	10292	286366	403
favourites_count	*integer*	User's favorites count	21621.4	0	459721	8713417	403
statuses_count	*integer*	User's statuses count	20551.9	8	246080	8282435	403
covid	*integer*	Number of tweets about covid	1367.8	1	6366	551208	403
politics	*integer*	Number of tweets about politics	1282.3	1	6180	516757	403
events	*integer*	Number of tweets about dissaster events	1360.0	1	6391	548099	403
emotions	*integer*	Number of tweets about emotions	1279.7	1	6173	515699	403
total	*integer*	Total Tweets considered	5289.7	4	25110	2131763	403
Sentiment	*integer*	Total Sentiment Emotion of each user	78.0	-577.5	660.0	31440.5	403
magnitude	*integer*	Total Sentiment Magnitude of each user	194.8	0.1	1053.1	78518.9	403
TweetRT	*integer*	Total Retweet count in sample of tweets of each user	3643993.2	0	59438254	1468529260	403
TweetFav	*integer*	Total Favorites count in sample of tweets of each user	2644.0	0	621342	1065515	403

TABLE VI. EXAMPLE OF SENTIMENT ANALYSIS FROM A USER

Tweet	**Score**	**Magnitude**	**TweetRT**	**TweetFav**
atrevemos corta vida	0.7	0.7	4072	0
imaginas valoren	0.2	0.2	4761	0
Que vida	0.6	0.6	0	0
Lo dice se retwittea	0	0	4942	0
imaginas sale bien	0.9	0.9	354	0
forever happy	0.9	0.9	3	0
Podria ser menos complicada entonces seria	0	0	8237	0
peores feel sentir quieres alguien que vano	-0.6	0.6	2	0

Appendix B. Codes Used in the Analysis