Revista Perspectiva Empresarial
ISSN: 2389-8186
perspectivaempresarial@ceipa.edu.co
Fundación Universitaria CEIPA
Colombia

SUÁREZ-GUTIÉRREZ, MANUEL; SÁNCHEZ-CERVANTES, JOSÉ LUIS; PAREDES-VALVERDE, MARIO ANDRÉS

An Online Social Network model through Twitter to build a social perception variable to measure the violence in Mexico

Revista Perspectiva Empresarial, vol. 7, núm. 2-1, julio-diciembre, 2020, pp. 6-18
Fundación Universitaria CEIPA

ISSN 2389-8186
E-ISSN 2389-8194

Vol.7, No. 2-1
Julio-diciembre de 2020
doi: https://doi.org/10.16967/23898186.665

ARTÍCULOS ORIGINALES

An Online Social Network model through Twitter to build a social perception variable to measure the violence in Mexico

pp. 6-18

MANUEL SUÁREZ-GUTIÉRREZ*
JOSÉ LUIS SÁNCHEZ-CERVANTES**
MARIO ANDRÉS PAREDES-VALVERDE***

* PhD Engineering in Emerging Technologies. Universidad Veracruzana, Veracruz, México. E-mail: mansuarez@uv.mx. ORCID: 0000-0003-0261-5765. Google Scholar: https://scholar.google.com/citations?hl=es&user=CtmAuecAAAAJ.

** PhD in Artificial Intelligence. Instituto Tecnológico de Orizaba, Veracruz, México. E-mail: jlsanchez@conacyt.mx. ORCID: 0000-0001-5194-1263. Google Scholar: https://scholar.google.com/citations?hl=es&user=pqnWv2AAAAAJ.

*** PhD in Computer Science. Instituto Tecnológico Superior de Teziutlán, Teziutlán, México. E-mail: marioandres.paredes@live.itsteziutlan.edu.mx. ORCID: 0000-0001-9508-9818. Google Scholar: https://scholar.google.com.mx/citations?user=AYJZ7cEAAAAJ&hl=es.

ISSN 2389-8186
E-ISSN 2389-8194

Vol.7, No. 2-1
Julio-diciembre de 2020
doi: https://doi.org/10.16967/23898186.665

COMO CITAR ESTE ARTÍCULO
How to cite this article:

Suárez-Gutiérrez, M., Sánchez-Cervantes, J.L. and Paredes-Valverde, M.A. (2020). An Online Social Network model through Twitter to build a social perception variable to measure the violence in Mexico. *Revista Perspectiva Empresarial*, *7*(2-1), 6-18.

Recibido: 20 de agosto de 2020
Aceptado: 07 de diciembre de 2020

ABSTRACT This paper describes the methodology and the model that used in Twitter to create an indicator that allows us to denote a social perception about violence, a topic of high impact in Mexico. We investigated and validated the keywords that Mexicans used related to this topic, in a specific time-lapse defined by the researchers. We implemented two analysis levels, the first one relative to the sum of tweets, and the second one with a rate of total tweets per 100,000 inhabitants. The results are geographically delimited, using a state and metropolitan zones scale.

KEY WORDS Social media, data analysis, pattern recognition, violence, perception, semantics.

Un modelo de red social en línea a través de Twitter para construir una variable de percepción social para medir la violencia en México

RESUMEN Este trabajo describe la metodología y el modelo utilizado en Twitter para crear un indicador que permita denotar una percepción social sobre la violencia, un tema de alto impacto en México. Se investigaron y validaron las palabras clave que los mexicanos utilizaron en relación con este tema en un lapso de tiempo específico definido por los investigadores. Se implementaron dos niveles de análisis, el primero relativo a la suma de tweets y el segundo con una tasa de tweets totales por cada 100000 habitantes. Los resultados se delimitan geográficamente, utilizando una escala de zonas estatales y metropolitanas.

PALABRAS CLAVE redes sociales, análisis de datos, reconocimiento de patrones, violencia, percepción, semántica.

Um modelo de rede social em linha através de Twitter para construir uma variável de percepção social para medir a violência no México

RESUMO Este trabalho descreve a metodologia e o modelo utilizado em Twitter para criar um indicador que permita denotar uma percepção social sobre a violência, um assunto de alto impacto no México. Se investigaram e validaram as palavras chaves que os mexicanos utilizaram em relação com este assunto num lapso de tempo específico definido pelos investigadores. Se implementaram dois níveis de análise, o primeiro relativo à soma de tweets e o segundo com uma taxa de tweets totais por cada 100.000 habitantes. Os resultados se delimitam geograficamente, utilizando uma escala de zonas estatais e metropolitanas.

PALAVRAS CHAVE redes sociais, análise de dados, reconhecimento de padrões, violência, percepção, semântica.

Introduction

The increase in the use of the Online Social Networks —OSN—, also known as Social Media, articulated with the growth of the Internet of Things —IoT—, where each of the electronic devices that we use (wireless sensors, GPS, smartphones, smartwatches, control systems) are connected to the Internet (Nguyen and Jung, 2018), creates a new space for generating and capturing large-scale data (Big Data) of any activity carried out by humans and intelligent machines. Social media has generated interest in vast proportions, especially in young people (users under 24 years). According to the Mexican Internet Association, in 2019, these users represent 44 % of the total number of users in Mexico. Additionally, they mention that the primary purpose of the connection is the use of OSN, with 82 % activity, where the Twitter site is ranked fifth nationwide popularity with an uptake of 39 % of internet users. The data generated and obtained from the OSN sites can be used for multiple purposes, such as prediction, digital marketing, analysis of sentiment, and even manipulation of social masses with fake news.

We focused on phenomena related to the interaction of the users on Twitter. These interactions allow to carry out studies like geolocated data (Brogueira, Batista and Carvalho, 2016), collaborative education (Kim, Hwang and Rho, 2016), sentiment analysis (Baydogan and Alatas, 2018), food industries (Singh, Shukla and Mishra, 2018) characterization of users (Lee, Wakamiya and Sumiya, 2013), disaster response (Xie and Yang, n.d.), drug war (Monroy-Hernández, Kiciman and Counts, 2015) and more. In this paper, we examine Twitter as the primary source to try to understand the social perception of violence in Mexico. The metadata analyzed allows establishing the origin of the tweet, identifying keywords, validating the information, as well as other data of interest. A web mining technique is necessary to build, determine, and verify the keywords. To do this, we used a Twitter analytics tool called TweetReach. Our contribution applies techniques of data analysis and modeling based on Big Data, via tools from the Elastic package (Elasticsearch, Logstash, and Kibana) to measure the degree of social perception of violence based on the publications on Twitter in Mexico, at a state and metropolitan zones scale.

This paper consists of five sections: (i) the introduction; (ii) a review of related works on perception analysis through the Twitter site, and the subject of violence in social media; (iii) the design of the methodology; (iv) the results are presented at a level of analysis of hashtags, and the degree of social perception of violence at a scale of federal entities and metropolitan areas; (v) finally, we presented our conclusions and remarks about future work.

Related Works

Perception analysis through Twitter has become a benchmark for identifying analysis patterns. The Twitter site has become a giant database of textual information on various topics like medicine (Devraj and Chary, 2015; Mahata et al., 2018), sentiment analysis (Bustos López et al., 2018; Garg, Garg and Ranga, 2017; Patankar, Kshama and Kotrappa, 2016; Salas-Zárate et al., 2020; Singh, Shukla and Mishra, 2018), digital marketing (Al-Hajjar and Syed, 2015; Nahili and Rezeg, 2018), to name a few.

This has been motivated by industry demands and research interest, where classifying text sentiment has become increasingly viable in the past decade.

Social Media, Twitter, and Big Data can help identify indicators about violence. Multiple authors consider different impact domains and techniques (Table 1). A description of citizens affected by the drug war in Mexico to measure the complex tensions between users that interact on the OSN was presented in (Monroy-Hernández, Kiciman and Counts, 2015). An innovative model based on the #MeToo movement to classify sexual assaults was exposed in (Khatua, Cambria and Khatua, 2018). A model to analyze students' stress level after violent events in a university campus was studied in (Saha and De Choudhury, 2017). A model to expose how right-wing politics is less tolerant over specific topics was revealed in (Ottoni et al., 2018). A correlation of violence and crime near stadiums in England influenced by tweets was analyzed (Ristea, Langford and Leitner, 2017).

Table 1. Selection of related works utilizing the subject of violence in social media

Research	OSN	Impact domain	Technique
Monroy-Hernández, Kiciman and Counts, 2015	Twitter	Drug War	Twitter data analytics
Khatua, Cambria and Khatua, 2018	Twitter	Sexual violence	Deep Learning
Saha and De Choudhury, 2017	Reddit	Stress on violence events	Machine Learning
Ottoni et al., 2018	YouTube	Violence aggression from videos	Semantic analysis, Natural Language
Ristea, Langford and Leitner, 2017	Twitter	Mass events violence	Pearson's correlation

Source: author own elaboration.

In summary, information coming from the Twitter site can be relevant in many domains related to violence, as we see in Table 1. Moreover, perception analysis through Twitter is an area of opportunity to establish preventive activities to mitigate or decrease the vulnerable sensation of security. Furthermore, the related works consulted, help us to deduce the importance of doing a correlation of tweets and other indicators such as social, economic, educational, safety, among others. Our contribution describes the procedure carried out, which can be replicated by changing the keywords of analysis for other studies.

Methodology

To carry out this study, we considered four issues: data capture and collection, storage, data analysis, and server availability. The proposed procedure to address these issues consists of the following steps: (i) selection of keywords; (ii) validation of keywords; (iii) configuration of the server cluster; (iv) data preprocessing; (v) data processing; (vi) elaboration of graphs and maps (results).

Selection of keywords

In the case of the social network of Twitter, in order to collect data, first, we defined and listed keywords (including hashtags and key user accounts). To get permission to the Twitter API, we requested authorization from Twitter to obtain the access keys. This API allows receiving only 1 % of the whole public dataset.

Then, we extracted keywords from tweets posted by user accounts from government institutions, where they published violence related events. These allowed verifying the profile of the user account that posted that information. We selected 64 Twitter user accounts linked to the government in all his levels (federal, state, municipal).

For keyword extraction, we used TweetReach. With this tool each user account selected is analyzed. TweetReach gives us a report with the top keywords published by the user account. So, we used the most published keywords by these selected users, giving us a total of 150 keywords.

Validation of keywords

For validating the list of keywords, we also used the TweetReach tool, which generates Twitter Analytics reports. These reports show the importance and relevance of the keywords based on the indicators of estimated reach and exposure, using a random sample of 100 tweets from the past few days. In Table 2, we show the five keywords with the most importance as a simple.

Table 2. Importance and relevance of the keywords

Keyword	Estimated Reach	Exposure	Importance of the keyword
Prevención del delito	1,074,307	3,838,922	72.02 %
Aprehendimos	1,514,997	5,242,496	71.10 %
Amenazada	729,112	2,226,465	67.25 %
Amagó	738,810	1,607,486	54.04 %
Sicarios	2,507,355	5,328,253	52.94 %

Source: author own elaboration.

Configuration of the server cluster

The configuration of a server cluster is implemented. The server cluster consists of four nodes, three of them were configured virtually on the Google Cloud™ platform, and one locally (Figure 1). Node 1 corresponds to the central server of the cluster, it stores the database being the central nucleus, and has Elasticsearch (Kononenko et al., 2014) installed. Node 2 visualizes the data stored in node 1 through the Kibana server. Node 3 manages the connection between the cluster and the Twitter API through Logstash (Langi et al., 2016) using a file in JSON format. This file contains three sections: input (containing access codes, list of keywords, geographic location, and language), filter, and output (establishing the connection to node 1, and indicating the output format). Finally, node 4 is the local server that performs the remote administration of the cluster, stores the backup of the database through 'Elasticdump,' downloads the DB, and uses "Oracle® MySQL" to manage the DB. Furthermore, this node must have the necessary tools to carry out the BD analysis.

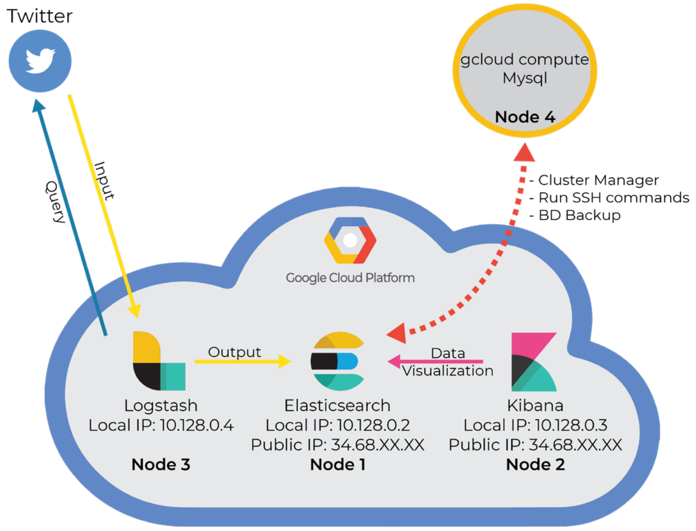


Figure 1. Server cluster configuration. Source: author own elaboration.

Data preprocessing

The data preprocessing phase allowed identifying those tweets that gathered the desired requirements for data analysis, as well as standardizing and normalizing the database of tweets so that it can be comparable. It was vital to remember that users write each tweet, and they have their writing peculiarities. They use special characters and emoticons, in addition to also restricting their vocabulary because of the limiting number of characters allowed on Twitter. The summary of the phases addressed in data preprocessing is shown next, as well as the main results obtained in each one of them:

1. *Data crawling* had a duration of 45 days, from May 14 to June 27 of 2019. During this period, we collected 20,736,887 tweets giving an average of 460,819 tweets per day.

2. *Data cleaning* consisted of excluding those misclassified tweets. For this, we verified the language of the tweet, and that the place of origin was Mexico. In other words, we excluded 2,983,421 (14.43 %) from the total tweets captured because they were not in Spanish. Nevertheless, we verified those tweets with the place of origin, identifying that they came from other countries like the United States, Brazil, Portugal, among others. To identify the place of origin of the tweets, we applied a total of 32 filters (which correspond to each of the Mexican states), excluding 17,753,466 (85.61 %) tweets for not coming from any Mexican state, leaving 983,336 tweets. Mostly of the excluded tweets came from homonymous words or cities. From this universe, we applied a filter to determine the total number of tweets with geographic information at the municipal level, leaving 507,023 tweets.

3. *Data normalization* consisted of standardizing the names of federal entities and each of the municipalities at the national level. To carry out this phase of standardization, we used the Unique Catalog of State, Municipal and Local Geostatistical Areas, published by the National Institute of Statistics and Geography (INEGI, for its Spanish acronym). Also, we made a word correction in the text metadata where we removed symbols, special characters, duplicate blank spaces, line breaks, and emoticons. We deleted a total of 42,249 emoticons and 2,944,765 special characters.

4. *Data validation* is related to the integrity of the datasets acquired from the Twitter site. However, some authors mention different techniques. Crowdsourcing semantics techniques to categorize the data as positive, negative, and neutral was used (Agarwal, Ravikumar and Saha, 2017). The divide and conquer method, graph-based modeling, and parallel data processing during data capture to improve the certainty and integrity of the data was applied (Senapati, Njilla and Rao, 2019). For the purpose of this study we applied the veracity index technique (a combination of geographic spread index, spam rate and diffusion index) (Ashwin, Kammarpally and George, 2016). We consider the variables (Ashwin, Kammarpally and George, 2016) to generating the veracity degree (VD) of a set of tweets, from combining indicators like dissemination index (ID), geographical extension index (IEG), and the relevant tweets index (ITR):

(i) Dissemination Index (ID): Corresponds to identifying the disseminating information speed on Twitter on issues of violence and insecurity in Mexico:

$$ID = 1 - \left(\frac{Total\ Unique\ Users}{Total\ Tweets}\right) = 1 - \left(\frac{94{,}960}{507{,}023}\right) = 0.8127$$

(ii) Geographical Extension Index (IEG): Identifies the dissemination of information on average to federal entities at a municipal level. Where the total locations are 2,457 municipalities spread over 32 states, there are also 74 Metropolitan Zones, which include 417 municipals:

$$IEG = \frac{\left(\frac{\sum State\ Reached}{Total\ States} + \frac{\sum Mun\ Reached}{Total\ Mun} + \frac{\sum Mun\ covered\ in\ ZM}{Total\ Mun\ of\ ZM}\right)}{3}$$

$$IEG = \frac{\left(\frac{32}{32} + \frac{1{,}121}{2{,}457} + \frac{312}{417}\right)}{3} = 0.7348$$

(iii) Relevant Tweets Index (ITR): Shows the importance of those unique tweets over the total number of tweets spread on the social network. In the case of this indicator, the closer the index is to 0, the more repeated tweets; therefore, there is a significant number of tweets considered as spam.

$$ITR = \frac{\sum unique\ tweets}{Total\ Tweets} = \frac{217{,}853}{507{,}023} = 0.4296$$

(iv) Veracity Degree (DV): It weighs the diffusion rates, geographic extension, and relevant tweets. In other words, it generates a weighted average of the three indicators, giving certainty about the dataset acquired through the independence of the opinions, scope, and impact of each of the Twitter users.

$$DV = \frac{(ID + IEG + ITR)}{3} = \frac{(0.8127 + 0.7348 + 0.4296)}{3} = 0.6590$$

The ID shows a value of 0.8127, indicating that users are very active as soon as an impacting news item related to the topic of violence occurs. The IEG has a value of 0.7348, derived from having 100 % coverage of the Federal Entities, 45 % of the Municipalities, and 75 % of the Municipalities that belong to a Metropolitan Area. The ITR shows a value of 0.4296, where many users forward a tweet that they perceive to be necessary. Finally, from the conjunction of these three indicators, the DV is obtained, reaching a value of 0.6590, validating the set of acquired tweets since it is more significant than 0.5, manifesting a higher correlation between the three indicators.

Data processing

The descriptive statistical analysis of frequencies classified two sections: the first corresponds to the set of data from a geographic scale at the level of states; the second corresponds to the data set at a geographical level of Metropolitan Zones. This in Mexico are established and delimited by the National Population Council (CONAPO, for its Spanish acronym), who mentions that in 2015, there were 74 Metropolitan Zones, which covered 17 % of the total municipalities (417 municipalities out of 2,456), and these agglomerated 63 % of the population nationwide (75,082,458 inhabitants out of the 119,530,753). According to the information shown in Table 3, the data classification has two indicators: the first related to the total number of tweets, and the second concerning the rate per 100,000 inhabitants.

(i) Total Tweets: The classification levels by quartiles, where we take the value of Q1 (7,193) as the first cut-off point, with four cut-off numbers and a width of 5,238, generate five ranges shown in Table 4.

(ii) Rate per 100,000 inhabitants: The classification levels by quartiles, taking the value of Q1 (280,733) as the reference for the first cut-off point, with four cut-off numbers and a width of 169,323, generate five ranges shown in Table 4.

Table 3. Descriptive statistical analysis by frequency for state level

Quartiles	Total Tweets	Rate per 100,000 inhabitants
Q1 (25)	7,193	280.733
Q2 (50)	11,673	352.391
Q3 (75)	17,250	547.226

Source: author own elaboration.

Table 4. Classification range of analysis intervals for state level

Total Tweets	Rate per 100,000 inhabitants	Rate
1,955-7,193	111.410-280.733	Very Low
7,194-12,431	280.734-450.056	Low
12,432-17,669	450.057-619.379	Medium
17,670-22,907	619.380-788.702	High
22,908, and more	788.703, and more	Very High

Source: author own elaboration.

In the case of descriptive statistical analysis at the Metropolitan Zones level (Table 5), we also considered two moments: the first refers to the total number of tweets, and the second refers to the rate per 100,000 inhabitants.

(i) Total Tweets in the Metropolitan Areas: we focused on the classification levels by quartiles, where we selected the value of Q1 (1,021) as the first cut-off point, with four cut-off numbers and a width of 1,020, generate five ranges shown in Table 6.

(ii) Rate per 100,000 inhabitants in the Metropolitan Zones: the classification by quartiles, taking the value of Q1 (254.8325) as the first cut-off point, with four cut-off numbers and width of 254,0025, generate five ranges shown in Table 6.

Table 5. Descriptive statistical analysis by frequency for Metropolitan Zones

Quartiles	Total Tweets	Rate per 100,000 inhabitants
Q1 (25)	1,021.25	254.8325
Q2 (50)	2,487.00	544.0500
Q3 (75)	7,323.00	760.9800

Source: author own elaboration.

Table 6. Classification of analysis intervals for Metropolitan Zones

Total Tweets	Rate per 100,000 inhabitants	Rate
1-1,021	0.8300-254.8325	Very Low
1,022-2,041	254.8326-508.8350	Low
2,042-3,061	508.8351-762.8375	Medium
3,062-4,081	762.8375-1,026.8400	High
4,082, and more	1,026.8401, and more	Very High

Source: author own elaboration.

Results

Hashtag analysis

Performing an analysis based on the hashtags of the tweets stored in the database, we determined that there are 8,600 different and unique ones, which are spread in 78,871 tweets, and by 23,324 users. The top 50 hashtags aggregate 24,773 tweets (Figure 2), representing 31.40 % of total tweets. The hashtag "#DíaNaranja" (Orange Day in Spanish) is the most widely used keyword. According to the National Commission to Prevent and Eradicate Violence Against Women (CONAVIM, for its Spanish acronym), the orange day is commemorated on the 25th of each month with the purpose of "acting, raising awareness, and preventing violence against women and girls."

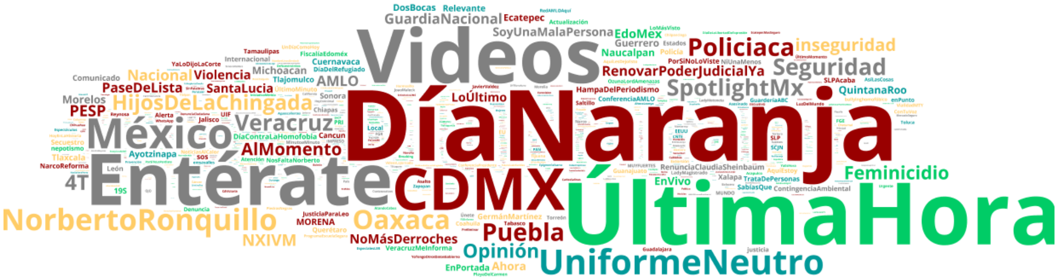


Figure 2. Word Cloud for most used hashtags. Source: author own elaboration.

We proposed a graphical representation of the knowledge acquired from the queries made on Twitter in graph format. This proposal allows us to understand the analysis of online social networks, by interweaving an algebraic, numerical model, with a taxonomic study of hashtags. In other words, we measure the frequency in which a term appears in each tweet to obtain the weights of each graph node, where a node corresponds to the published hashtags.

On the other hand, the edges correspond to the interaction of the nodes among them. To achieve this, firstly, we made a query in the database, obtaining the list of hashtags with the users who published them (that is, the origin of the edge). Secondly, we selected the group of users identified by the hashtags that they disseminated (edge destiny). Also, we deleted hashtags that point to themselves (circular edges). The graph shows that Figure 3, has 1,343 nodes and 152,316 edges, representing all those tweets that have a degree weight attraction between them. Besides, it allows recognizing the existence of hashtags or nodes with greater relevance and importance based on the size of the node, which is proportional to the number of times users cited them. Similarly, we identified edges by color and line thickness (weight), that is, the number of times that a node is related to another node.

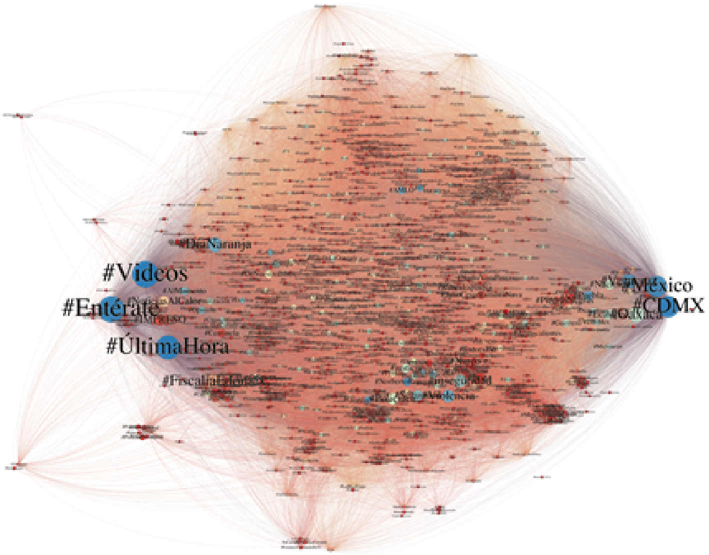


Figure 3. Hashtag correlation. Source: author own elaboration.

State-level analysis

The indicator *total tweets grade* classifies them according to the number of tweets sent by all entities, establishing five intervals (Very Low, Low, Medium, High, and Very High). We obtained the following results for the 32 states: 6 with a Very High grade rating, representing 49 %; 1 with a High grade, concentrating 5 %; 7 with a Medium grade, indicating 21 %; 10 entities with a Low grade, signifying 19 %; 8 with a Very Low grade, showing 6 %. Therefore, the higher the grade of the indicator for the total number of tweets, the higher publication of tweets that come from that entity.

In Figure 4, we show the results corresponding to the grade of total tweets. The entities in red (Veracruz, Ciudad de México, Jalisco, Nuevo León, and Sonora) represent a very high index. The rate per 100,000 inhabitants is also classified into five intervals (Very Low, Low, Medium, High, and Very High). Derived from this classification, we obtained the following results for the 32 states: 3 with Very High; 3 with High; 7 with Medium; 11 with Low; 8 with Very Low. Consequently, the higher the rate per 100,000 inhabitants, the more significant is the perception of violence, based on the statements posted on the Twitter site by the total population for each entity (Figure 4). Regarding the rate per 100,000 inhabitants, it represents the correlation of the total tweets and the residents of the states. We obtained that the entity with the highest participation of events in the topic of violence on the Twitter site corresponds to: Nuevo León with 1,141 tweets; Quintana Roo with 1,074 tweets; Sonora with 814 tweets; Morelos with 716 tweets; and Jalisco with 696 tweets.

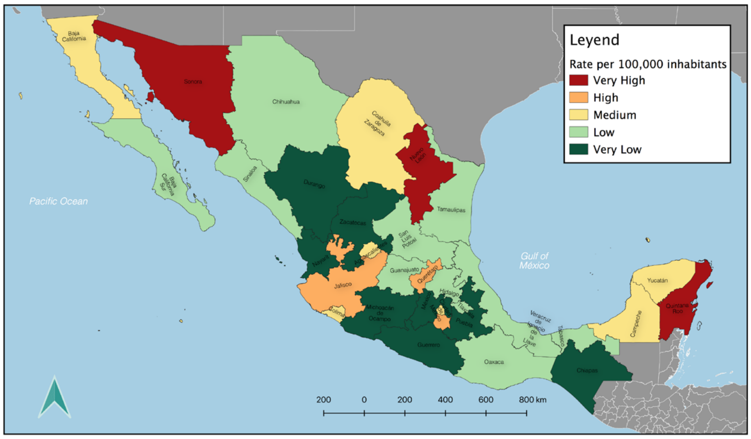
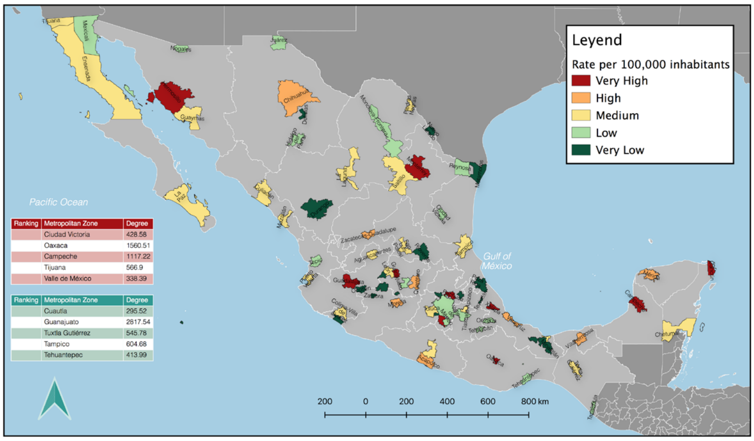


Figure 4. Map of the total tweets degree and rate per 100,000 inhabitants at a state level. Source: author own elaboration.

Metropolitan Zone analysis

The indicator at a Metropolitan Zones scale (Figure 5) visualizes the degree of perception on the spread of acts of violence in specific geographic areas. We identified that 24 % had a Very High or High grade, 28 % with a Medium grade, and 47 % with a Low or Very Low grade. The data shown make possible the understanding and impact that Twitter has in each of the Metropolitan Areas. Also, 7 out of the 74 Metropolitan Zones (Valle de México, Monterrey, Guadalajara, Hermosillo, Toluca, Querétaro, and Puebla-Tlaxcala) generate 50 %of the total number of tweets captured, while 17 spread less than 1,000 tweets.

We detected that 29 Metropolitan Zones have a Very High grade, encompassing 86.91 % of the total tweets (403,142). Out of these, 9 (Guanajuato, Hermosillo, Oaxaca, Cancún, Xalapa-Enríquez, Monterrey, Cuernavaca, Guadalajara, and Pachuca) have a Very High grade regarding the rate per 100,000 inhabitants, this means that they publish a lot about violence, and they also feel vulnerable in their cities. On the other hand, zones like Valle de México and Puebla-Tlaxcala, with Very High grade of total tweets, have a Low grade on the rate per 100,000 inhabitants. It can be inferred that; they publish a lot of tweets (17.83 % of the total) because they have a high population density (31.74 % of the total in metropolitan zones).



Ranking	Metropolitan Zone	Degree
	Ciudad Victoria	428.58
	Oaxaca	1560.51
	Campeche	1117.22
	Tijuana	566.9
	Valle de México	338.39

Ranking	Metropolitan Zone	Degree
	Cuautla	295.52
	Guanajuato	2817.54
	Tuxtla Gutiérrez	545.78
	Tampico	604.68
	Tehuantepec	413.99

Figure 5. Map relative to the rate per 100,000 inhabitants at a metropolitan zone level. Source: author own elaboration.

Conclusions

This paper proposes the determination of a methodology that allows measuring the impact of the indicator of Mexican social perception of violence at a states and Metropolitan Zones scale, based on data collected on the online social network of Twitter during a period of time. The analysis was carried out with 507,023 geo-localized tweets at a state and municipal level in Mexico from May 14 to June 27, 2019.

The Twitter site can track data by measuring the degree of social perception of violence on a states and metropolitan areas scale in Mexico. With this indicator, we can detect socio-territorial inequalities, keeping in mind that we can only collect data in spaces with an Internet connection. With this approach, we can suggest and infer whether acts of violence can limit or motivate public and private investment in specific territorial spaces.

For future research, we need to work in a multidisciplinary way to contrast the result of the proposed indicators against other indicators such as social and economic, to name a few. The purpose is to understand the social environment to explain and justify the reason why an entity obtained an extraordinary or deficient degree of violence perception.

References

Agarwal, B., Ravikumar, A. and Saha, S. (2017). A Novel Approach to Big Data Veracity Using Crowdsourcing Techniques and Bayesian Predictors. In 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, USA.

Al-Hajjar, D. and Syed, A. (2015). Applying Sentiment and Emotion Analysis on Brand Tweets for Digital Marketing. In IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan.

Ashwin, K., Kammarpally, P. and George, K. (2016). Veracity of Information in Twitter Data: A Case Study. In International Conference on Big Data and Smart Computing (BigComp), Hong Kong, China.

Baydogan, C. and Alatas, B. (2018). Sentiment Analysis Using Konstanz Information Miner in Social Networks. In 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, Turkey.

Brogueira, G., Batista, F. and Carvalho, J.P. (2016). Using Geolocated Tweets for Characterization of Twitter in Portugal and the Portuguese Administrative Regions. *Social Network Analysis and Mining*, *6*(1), 1-20.

Bustos López, M. et al. (2018). EduRP: An Educational Resources Platform Based on Opinion Mining and Semantic Web. *Journal of Universal Computer Science*, *24*(11), 1515-1535.

Devraj, N. and Chary, M. (2015). How Do Twitter, Wikipedia, and Harrison's Principles of Medicine Describe Heart Attacks? In Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, Atlanta, Georgia.

Garg, P., Garg, H. and Ranga, V. (2017). Sentiment Analysis of the Uri Terror Attack Using Twitter. In International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India.

Khatua, A., Cambria, E. and Khatua, A. (2018). Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain.

Kim, Y., Hwang, E. and Rho, S. (2016). Twitter News-in-Education Platform for Social, Collaborative, and Flipped Learning. *The Journal of Supercomputing*, *74*(8), 3564-3582.

Kononenko, O. et al. (2014). Mining Modern Repositories with Elasticsearch. In Proceedings of the 11th Working Conference on Mining Software Repositories, Chicago, USA.

Langi, P. et al. (2016). An Evaluation of Twitter River and Logstash Performances as Elasticsearch Inputs for Social Media Analysis of Twitter. In International Conference on Information & Communication Technology and Systems (ICTS), Surabaya, Indonesia.

Lee, R., Wakamiya, S. and Sumiya, K. (2013). Urban Area Characterization Based on Crowd Behavioral Lifelogs over Twitter. *Personal and Ubiquitous Computing*, *17*(4), 605-620.

Mahata, D. et al. (2018). Detecting Personal Intake of Medicine from Twitter. *IEEE Intelligent Systems*, *33*(4), 87-95.

Monroy-Hernández, A., Kiciman, E. and Counts, S. (2015). Narcotweets: Social Media in Wartime. *Artificial Intelligence*, 515-518.

Nahili, W. and Rezeg, K. (2018). Digital Marketing with Social Media: What Twitter Says! In 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), Tebessa, Algeria.

Nguyen, H.-L. and Jung, J.E. (2018). SocioScope: A Framework for Understanding Internet of Social Knowledge. *Future Generation Computer Systems*, *83*, 358-365.

Ottoni, R. et al. (2018). Analyzing Right-Wing Youtube Channels: Hate, Violence and Discrimination. In Proceedings of the 10th ACM Conference on Web Science.

Patankar, A., Kshama, K. and Kotrappa, S. (2016). Emotweet: Sentiment Analysis Tool for Twitter. In EEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT), Pune, India.

Ristea, A., Langford, C. and Leitner, M. (2017). Relationships between Crime and Twitter Activity around Stadiums. In 25th International Conference on Geoinformatics, Buffalo, USA.

Salas-Zárate, M. et al. (2020). Review of English Literature on Figurative Language Applied to Social Networks. *Knowledge and Information Systems*, *62*(6), 2105-2137.

Saha, K. and De Choudhury, M. (2017). Modeling Stress with Social Media around Incidents of Gun Violence on College Campuses. In Proceedings of the ACM on Human-Computer Interaction.

Senapati, M., Njilla, L. and Rao, P. (2019). A Method for Scalable First-Order Rule Learning on Twitter Data. In IEEE 35th International Conference on Data Engineering Workshops (ICDEW), Macao, China.

Singh, A., Shukla, N. and Mishra, N. (2018). Social Media Data Analytics to Improve Supply Chain Management in Food Industries. *Transportation Research Part E: Logistics and Transportation Review*, *114*, 398-415.

Xie, J. and Yang, T. (n.d.). Using Social Media Data to Enhance Disaster Response and Community. In International Workshop on Big Geospatial Data and Data Science (BGDDS), Wuhan, China.