Analysing gender-based violence against Colombian public figures on Twitter

Juan Sebastian Chaparro-Saenz, Ixent Galpin

Facultad de Ciencias Naturales e Ingeniería, Universidad de Bogotá Jorge Tadeo Lozano, Bogotá, Colombia

Abstract

This work proposes the development of a methodology that standardises the extraction, processing and analysis of natural language data for the study of gender-based violence evidenced on the Twitter social network. We develop a tool that may be exploited by different organisations, foundations, corporations, associations or state institutions that promote, exercise and disseminate human rights in Colombia and elsewhere. In this work, we take as a case study ten prominent female public figures in Colombia in the artistic, political and journalistic spheres. We extract a total of 39,629 tweet responses during a turbulent national strike amid the COVID-19 pandemic, and carry out topic identification and sentiment analysis. While we observe differences between the different roles based on natural language processing with different libraries, the are notable negative terms in the topics identified which are of concern as they may incite gender-based violence. It is expected that this proposed tool will benefit the decision-making of these institutions to issue early warnings, together with the exercise of the protection, prevention and defence of women's rights.

Keywords

Sentiment Analysis, Topic Identification, Text Mining, Gender-based violence

1. Introduction

Violence against women has been highlighted as a problem that generates an impact of significant importance to society [1]. Such violence may take different forms, including physical, sexual, psychological, economic, verbal or written. These different types of violence can be exercised by different actors, including partners, colleagues, fellow students and even adversaries in the political field [2]. In Colombia, according to the *Defensoría del Pueblo* (Human Rights Ombudsman), there has been a increase in gender-based violence in the country since the start of the COVID-19 pandemic [3]. This includes different types of violence such as physical violence (18%), sexual violence (6%), psychological violence (42%), violence against assets (6%) and economic violence within a home (27%) [3].

Currently, society is constantly growing and evolving as part of the the digital era [4]. This in its wake allows different types of digital content to be created, giving way to a world in which it is possible to constantly interact, becoming routine for human beings. According to figures approximately 4.5 billion people currently use the Internet and 3.8 billion users are

ICAIW 2021: Workshops at the Fourth International Conference on Applied Informatics 2021, October 28–30, 2021, Buenos Aires, Argentina

juans.chaparros@utadeo.edu.co (J. S. Chaparro-Saenz); ixent@utadeo.edu.co (I. Galpin)
0000-0003-3741-5448 (J. S. Chaparro-Saenz); 0000-0001-7020-6328 (I. Galpin)

registered in a social network such as Facebook, Twitter, YouTube, WhatsApp, among others [5]. As such, this work focuses on the exploration of misogynistic postings on the Twitter social network. This is deemed to be of high importance as such negative tweets in turn trigger demonstrated psychological violence in Colombia, through intimidating comments, harassment, threats, contempt, mockery, humiliation, among others [3].

The microblogging Twitter platform is often considered to be a barometer of society [6], given that postings (or *tweets*) can be publically posted in real time. It enables people to express their opinions through publications maintaining freedoom of expression on any topic that is being debated at a moment in time. Twitter has become the *de facto* medium where opinions are expressed on different issues, especially in the political sphere, which usually generate contrasting and often conflicting points of view. On many occasions, these points of view reflect in gender violence [7]. In Colombia, Twitter currently has 3.2 million users, which accounts for approximately 7.8% of the population. While there is a significant digital divide in Colombia between rural and urban areas [8], there is also a gender-based one [9]. In the case of Twitter, this is also visible: 62.9% of users are male, compared to 37.1% female [10].

In this work, we select ten Colombian public figures in the political, artistic, or journalistic spheres. We collect and analyse *tweets* that correspond to responses to the ten public figures under study, based on the controversies that may arise from their opinions. We identify the predominant topics in these responses, and apply automated techniques to gauge sentiment (negative or positive) and the degree of subjectivity or objectivity. We apply the well-established CRISP-DM methodology [11] to ensure a robust analysis. This paper is structured as follows. Section 2 presents related work. Subsequently, we broadly follow the first five steps of the CRISP-DM methodology: Section 3 presents steps 1 and 2, business and data understanding. In Section 4, we describe the data processing carried out prior to analysis (step 3 of CRISP-DM). The next section describes the models applied. In Section 6, we present the results obtained through the evaluation of the models. Finally, we present a discussion in Section 7, and Section 8 concludes.

2. Related Work

At the time of writing, there is considerable research based on text mining applied to social problems such as violence, health, poverty, among others, that in turn supports decision-making.

For example, Cremades *et al.* [12] proposes the development of a system based on artificial intelligence capable of predicting a potential suicide. Such a system would be based on the application of a methodology that involves the collection, compilation and selection of text for processing. Saura *et al.* [13] analyse text from the `#BlackFriday` hashtag on Twitter. The authors conclude that companies do not use the social network as a marketing strategy since the study of this publication is based on the selection of opinions with a criterion related to exclusive offers. In summary, the study identifies the neutral, positive and negative sentiment on the consolidation of writings on a selection criterion [13].

The proposed methodology is integrated by stages for the final result. For this reason, it is necessary to employ data engineering approaches for the extraction and consolidation of the sample under study. Barriga *et al.* [14] employ technological elements for the extraction, consolidation and analysis of texts from the social network Twitter, such as the platform's own API that allows it to be integrated into the development of software under the Java language. Thus, it enables the use of methods for the extraction of tweets. Once the previous process occurs, there is a step involving warehousing of the information in two data models; relational and non-relational. A non-relational database engine, MongoDb, is used to contain the information of the extracted account profile and its corresponding timelines, treated as JSON files. Finally, the author aims to develop a web tool for the extraction and storage of data from the social network Twitter, which allow interoperability with external tools for data analysis [14].

Silva *et al.* [15] present an investigation directly related to social problems based on gender violence. In this case, technology in the location of some type of violence is applied, as it could be evidenced in verbal or written expression. Thus, a model for detecting aggressiveness towards women in opinions published on social networks was created through the application of machine learning techniques. It carries out the choice of a Twitter corpus, by means of extraction with the platform's API, performs a data cleaning process and makes use of Microsoft Excel for punctuation, grammar and spelling cleaning. Characteristics of the tweet are selected under a process developed in Java for further evaluation. Classification and evaluation of texts is carried out using a collection of machine learning algorithms with the Weka library. Finally, the result is the classification of a tweet as "aggressive" or "slightly aggressive" given the performance of each model evaluated.

Various studies analyse the behaviour of sentiments using Twitter hashtags. Evovli *et al.* [16] investigate, through the analysis of tweets, hatred which was directed towards Muslims following the United Kingdom Brexit referendum in 2016. A qualitative analysis is carried out of tweets with the hashtags `#IslamIsTheProblem` and `#Muslimterrorists`. In this study, variables that affect the study such as trolls and bots are taken into account. Evovli proposes that Islamophobia may be mitigated by considering the connections in the social network graphs [16].

3. Business and Data Understanding

To understand the problem, it is appropriate to understand the type of abusive and specific language towards women that will lead to the development of the tool in this work. This type of violent language directed at women is known as *misogyny*, defined as the hatred or prejudice towards women as a result of a belief that they are the weaker sex. Pamungkas *et al.* [17] detected the presence of misogyny through a series of cross-lingual classification experiments.

The empowerment of women as protagonists in the different roles of society has been noteworthy in recent years. The use of social networks has been a means of promoting ideals, relating and acquiring visibility. For this reason, we see the need to standardise a methodology for the analysis of gender violence for a specific group of women. We select a group who not only has political participation in the country, but are also an interdisciplinary group in order to address and characterise through the study the different types of language expressed in the opinions expressed about the opinions of these women.

For the proposed objetive, it is necessary to consolidate a representative sample of data according to the criteria that have been proposed. Thus, as a first step, a data engineering

Table 1
Selected profiles, number of tweet responses collected, and profile description

Name	Number of Tweet Responses	Profile
Claudia Lopez	5296	Mayoress of Bogotá 2020-2023
Martha Lucia Ramirez	4468	Vice President of Colombia 2018-2022
Margarita Rosa de Francisco	4468	Colombian actress, singer, writer, composer and presenter
Vicky Davila	5044	Colombian journalist
Maria Fernanda Carrascal	1900	Colombian activist
Angelica Lozano	5154	Senator of the Republic of Colombia
Paloma Valencia	5529	Senator of the Republic of Colombia
Claudia Gurissati	1440	Colombian journalist
Adriana Lucia	3468	Colombian singer and songwriter
Angela Robledo	2862	Senator of the Republic of Colombia

process is designed for the extraction of tweets through the implementation of the application programming interface of the Twitter platform. In this way, a connection point is generated between the program developed under the Python language that understands interoperability with the social network, concerning a specific type of information by users who decided to share it publicly. Subsequently, we aim to understand the degree of objectivity of opinion given the sentiment resulting from the adoption of natural language processing on the data source made up of ten female gender profiles. We select profiles with a considerable number of followers and constitute well-known public figures.

The sample selected, shown in Table 1, comprises activists, political leaders, journalists, actors, and singers within Colombian who have a significant Twitter presence. Through their opinions about political issues, they have managed to generate widespread impact and dissemination of their opinions, resulting in controversy and mixed reactions by other Twitter users. Each one of these women registers many reactions and replies to the tweets that they publish. This leads to an an increase in followers and a greater number of interactions, thus allowing an extensive source of data to be collected for the analysis of this study. The data sample was collected in the period between May and August 2021, a period where in Colombia economic activities were paralysed as a result of of roadblocks and protests, aggravated by the economic hardship arising the COVID-19 pandemic. In total, 39,629 tweet responses are collected, an average of 3,963 per profile.

4. Data Preparation

Taking into account the data collected from the profiles listed in Table 1, the development of a tool is carried out using Python[1], a scripting language commonly used for data science projects. Python as a language has become a well-established tool for data analysis and the main advantage is that it can be used without licensing costs. In other words, it is an open and free

[1] https://www.python.org/

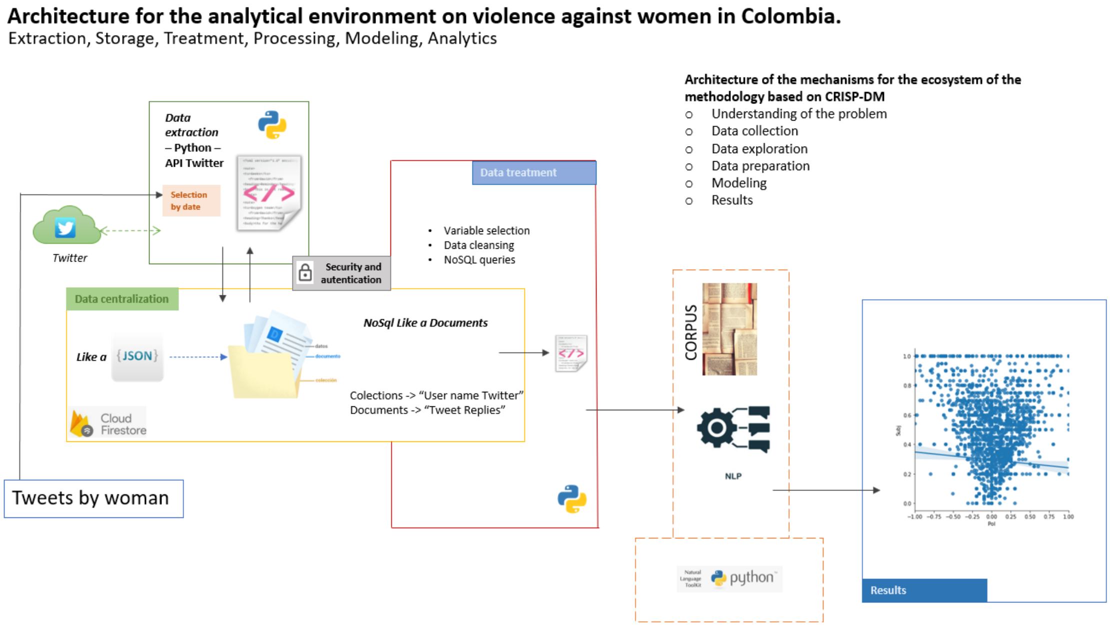

Architecture for the analytical environment on violence against women in Colombia.
Extraction, Storage, Treatment, Processing, Modeling, Analytics
Data extraction – Python – API Twitter
Selection by date
Twitter
Security and autentication
Data treatment
Variable selection
Data cleansing
NoSQL queries
Data centralization
Like a {JSON}
NoSql Like a Documents
Colections -> "User name Twitter"
Documents -> "Tweet Replies"
Cloud Firestore
Tweets by woman
Architecture of the mechanisms for the ecosystem of the methodology based on CRISP-DM
Understanding of the problem
Data collection
Data exploration
Data preparation
Modeling
Results
CORPUS
NLP
python
Results

Figure 1: Architecture for the analytical environment on violence against women in Colombia.

technology compared to proprietary technologies. As shown in Figure 1, the tool comprises four phases that govern the design and reflect critical aspects in accordance with the methodology adopted, viz.

1. Tweet extraction
2. Data cleaning
3. NoSQL database storage on Firestore Cloud
4. Text preprocessing.

4.1. Tweet extraction

In this phase, version 2 of the Twitter application platform programming interface is adopted[2] to search for tweets with specific criteria. It was configured to find all that tweets that reply to tweets created by the profiles listed in Table 1.

4.2. Data cleaning

Once the information was extracted, a treatment was carried out to eliminate repeated tweets, tweets written by the account owner and false tweets, in order to clean and build a valid data set for further study.

[2] https://developer.twitter.com/en/docs/twitter-api/early-access

4.3. NoSQL database storage on Firestore Cloud

Based on the information processed and collected, it is necessary for this tool to make use of catalog services that allow the set of semistructured data to be managed in an agile and fast way. In such a way, it is sought that this technological service complies with some characteristics that interoperate within the proposed model that makes use of a development based on Python language and consolidates a unified and solid base for consulting a large number of data. For this reason, the Google Cloud Firestore[3] database was chosen, which is a document-oriented NoSQL engine. Since our data set is treated as a JSON file in this way, Firestore allowed us to store each data set in different collections that relate a set of tweets per profile. Each document contains a set of key-value pairs that uses few resources and contains fields with assigned values [18].

4.4. Text preprocessing

By virtue of the compression on the objective of this study, it is necessary to become familiar with the initial data collection in storage in the NoSQL service, verifying the quality and quantity of these. In this sense, and in accordance with the flow of the CRISP-DM methodology, it is essential to understand the problem that needs to be solved, which is why the different techniques for data pre-processing are executed. After the discovery and preparation stage, resulting elements derived from threads are obtained, such as data cleaning, tokenization, stopword removal, stemming and lemmatization, closely related for the topic and classification model [19]. For the application of these techniques the NLTK[4] and Spacy[5] libraries were employed, which provide the possibility of treating the process in Spanish. Thus, the text preprocessing proposed for the present study consists of:

- *Data cleaning*: In this step the elimination of characters, numbers, punctuations was carried out.
- *Tokenization*: In this step we convert sentences into a word list.
- *Stopword Removal*: A list of words that do not provide correct information to the model is constructed and therefore the removal of these words is carried out.
- *Lemmatisation*: This technique is used to reduce the dimensionality of a word, that is, to take the verb to its infinitive form. In addition, suffixes that derive in quantities of something are removed.
- *Stemming*: As with the lemmatisation step that seeks to reduce words, this technique is done by applying structure rules on a set of letters joined in a word.

5. Modeling

Within the framework of the process proposed for the scope of this modeling phase, we explored the topics and sentiments according to the corpus obtained from each Twitter profile in this study.

[3] https://firebase.google.com/docs

[4] https://www.nltk.org/

[5] https://spacy.io/

Table 2
Parameters for the LDA model

Parameter	Value
Number of topics	4
Size of the document looked at every pass	500
Number of passes through documents	20
Iterations	400
Evaluation of the perplexity of the model	1

5.1. Topic Analysis using Latent Dirichlet Allocation

To understand the structure of the adopted model that determines the topics, Latent Dirichlet Allocation (LDA) was used, which consists of a three-level hierarchical Bayesian model. According to Blei *et al.* [20], "Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document". We use the LDA Gensin Models library[6] for implementation. The corpus is partitioned for each profile, allowing a visualisation with the most relevant topics to be generated. Once we have treated the data in which we seek to reduce the dimension of the vocabulary for the optimisation of the model, subsequently it can be identified that some words do not have an adequate meaning in isolation. In comparison, it is more meaningful if they are considered with adjacent words. Thus, it is necessary to consider n-grams [21]. As such, the data set is explored and the frequency with which the words appear together is identified, classifying them in bi-grams or tri-grams. Once we obtain the processed and cleaned tweets, we can build a dictionary to continue the training of the LDA model. In this step, the corpus vocabulary is built in which all the unique words in the data set are assigned a unique ID.

The subsequent *model training* step includes the use of the dictionary corresponding to the scenario, to refer a class to each of the topics. During development, it will be necessary to iterate multiple times to return the topics resulting from the most probable words. In order to find the optimal parameters for the LDA model defined in the development of the tool, they were initially defined in Table 2. With these parameters it is possible to contrast the most relevant topics and the slightly more frequent words.

5.2. Sentiment analysis

We subsequently carry out a sentiment analysis over each tweet. We use the Vader[7] and Textblob[8] libraries, which enable the measurement of *polarity* (i.e., how negative or positive a tweet is), and *subjectivity* (i.e., whether a tweet corresponds to fact or opinion) associated with a tweet. In accordance with the capabilities of these libraries, they do not support the Spanish language for sentiment analysis. For this reason, a process for the translation of text

[6] https://radimrehurek.com/gensim/models/ldamodel.html

[7] https://pypi.org/project/vaderSentiment/

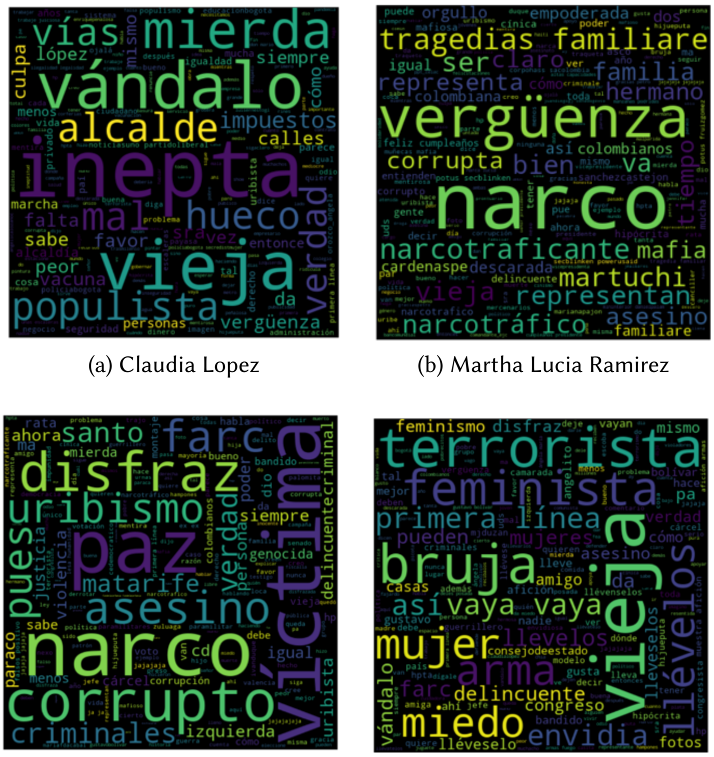
[8] https://textblob.readthedocs.io/en/dev/

was developed that consists of the installation of the “Translate Text” extension[9], which allows, without any limitation, to translate documents stored on Google Firestore Cloud into the English language.

6. Results

For the results chain, an orientation derived from mechanisms such as the word cloud, LDA gensim model, Vader sentiment library and Textblob library were considered to understand and identify the data that would allow discussions and insights on the violence evidenced on Twitter. The results obtained correspond to an estimate that would be obtained in a real scenario. The error of this estimate can be given by the concentration of elements that do not have an adequate meaning, the precision of the libraries, and the quality of the processed data. However, the method adopted constitutes an approximation of a real and a valid scenario to discuss violence against women.

6.1. Topic Analysis using Latent Dirichlet Allocation



(a) Claudia Lopez

(b) Martha Lucia Ramirez

(c) Paloma Valencia

(d) Angela Robledo

Figure 2: Word clouds generated for each profile

[9] https://firebase.google.com/products/extensions/firestore-translate-text

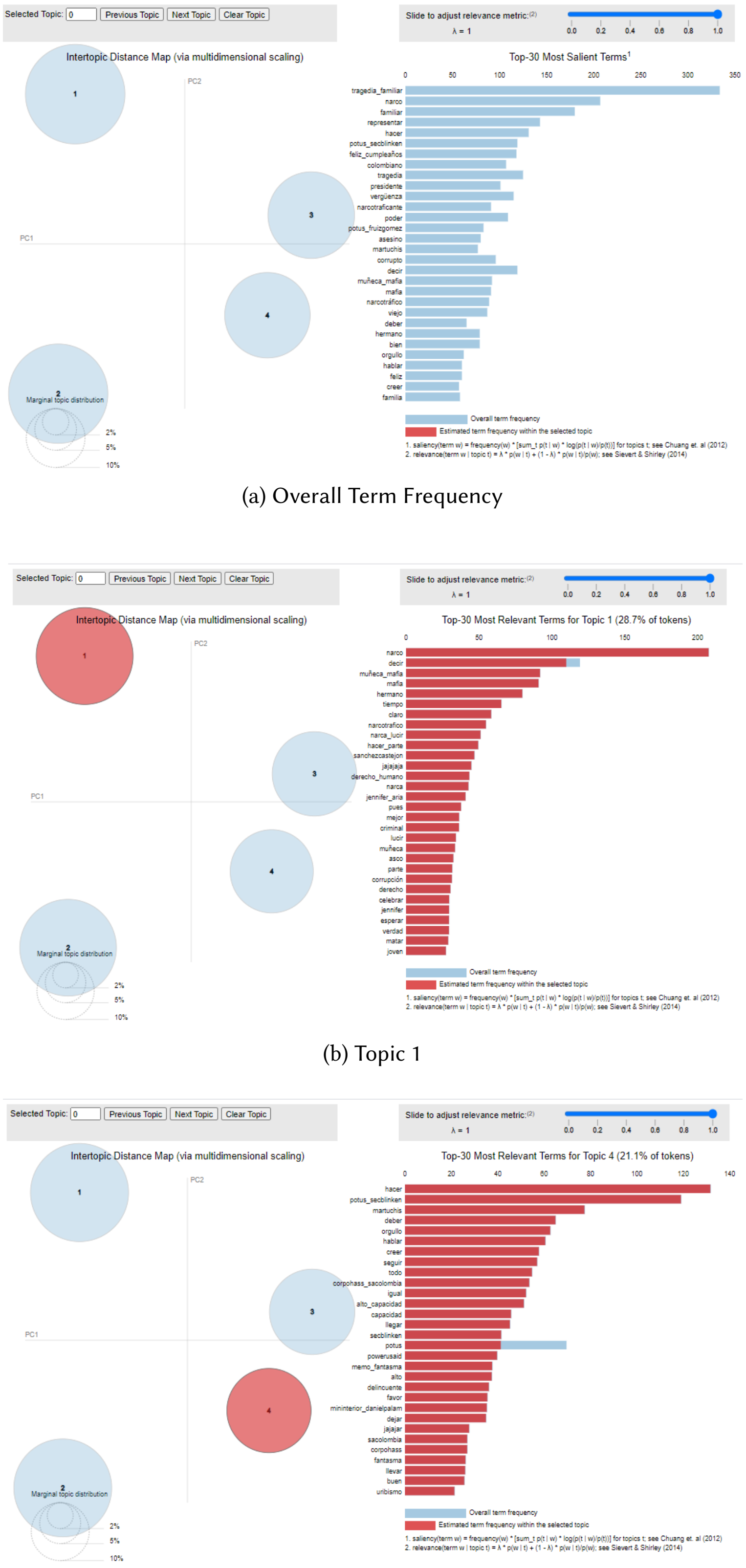
Table 3
Most salient negative terms

Name	Most salient negative terms
Claudia Lopez	vandalo,ineptar,mierda
Martha Lucia Ramirez	narco, muñeca_mafia, narcotráfico, narca_lucir, narca, muñeca
Margarita Rosa de Francisco	izquierda_destruir, definitivamente_tocar_izquierda_destruir
Vicky Davila	rata, bobo, uribista, doble_moral, asesinar, titere
Maria Fernanda Carrascal	robar
Angelica Lozano	angélico, muñeca jugadita, mierda
Paloma Valencia	criminal, victima, matarife, terrorista, hijueputa, delincuente
Claudia Gurissati	cobarde
Adriana Lucia	vandalo, maldad
Angela Robledo	terrorista, viejo, delincuente, vandalo, criminal

Figure 2 presents a graphical representation of the vocabulary as a visual resource of the most significant words within the data set obtained from the opinions whose content is integrated by the single user of the platform. Notably we see these cases with words in common between them such as "narco" and "vieja [a derogatory term used in Colombia to refer to a woman]", as well as outstanding words such as "mierda [shit]", "bruja [witch]", "inepta [inept]", all of them with a negative and violent connotation. These terms clearly suggest the presence of tweets with derogatory, discriminatory and stigmatising content. Furthermore, this finding reflects a social and cultural problem derived from history in which men have subjected women in multiple ways, developing dominance and stigmatising them as the weaker sex.

Figure 3 presents a visualisation with the identification of topics observed, with an adjustment of the λ parameter equal to 1. Following trial and error, four topics are found to yield the most meaningful characterisation of the data for each profile. This parameter will determine the weight given to the probability of a word on the topic. Now, as long as this adjustment is closer to 1, it will return a set of terms characterised by their probability in the topic. Each circle in the visualisation represents a topic, and the larger it is, the more dominant it will be compared to other topics. For the specific analysis of Martha Lucia Ramirez and Paloma Valencia, the most negative words are notably evident in comparison with the others, within the set of most outstanding terms for each topic, shown on the right hand side of each visualisation. In the case of the vice-president Martha Lucia Ramirez, we see that the term "narco" is the most prevalent within the major topic, and in turn is related to the second topic. Within the topics words are observed that suggest that the tweets contain grotesque, stigmatising and derogatory messages.

Table 3 presents the most salient negative terms for the topics identified in each profile, including n-grams with a political and slightly negative context. This finding allows us to show the presence of violent, derogatory and stigmatizing comments in this case in two profiles with a political roles. When observing the results for the singer Adriana Lucia, the absence of violent terms is observed, in comparison with the political roles where we observe a large index of elements that make up misogyny, that is, it is considered that in this area there is a violation of rights.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

(a) Overall Term Frequency

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

(b) Topic 1

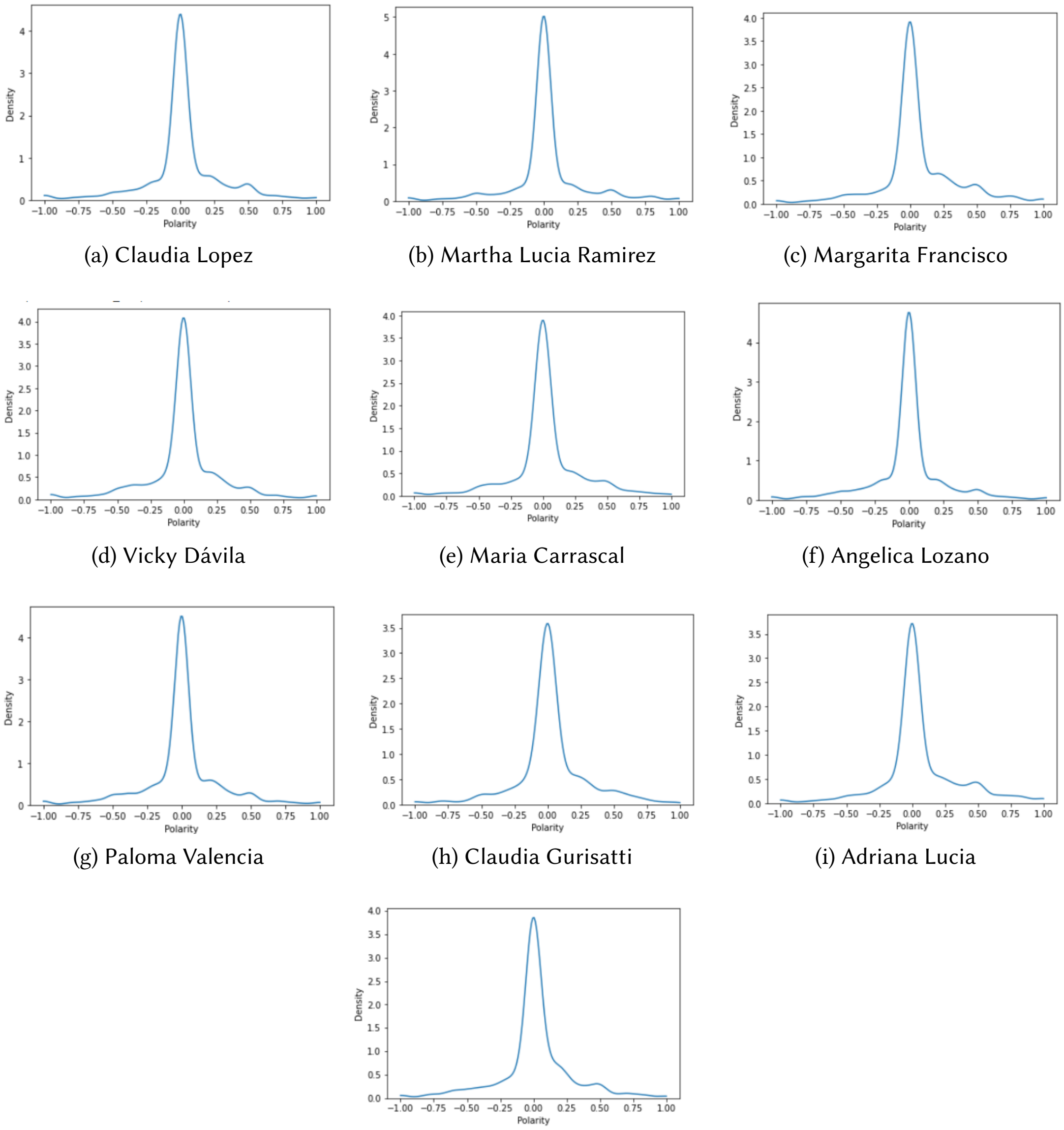
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

(c) Topic 4

Figure 3: Exploring Topics Using LDA Gensim for Martha Lucia Ramirez

6.2. Sentiment analysis

Figure 4 shows the distribution of polarity values for the tweets associated with each profile, in which -1 indicates a negative sentiment, 0 a neutral sentiment, and 1 a positive sentiment. It is observed that, by and large, tweets are associated with a neutral sentiment, for all the profiles in the study. Figure 5 shows the distribution of subjectivity values for the tweets associated with each profile, where 0 indicates a concrete and demonstrable fact, and 1 an opinion, emotion



(a) Claudia Lopez

(b) Martha Lucia Ramirez

(c) Margarita Francisco

(d) Vicky Dávila

(e) Maria Carrascal

(f) Angelica Lozano

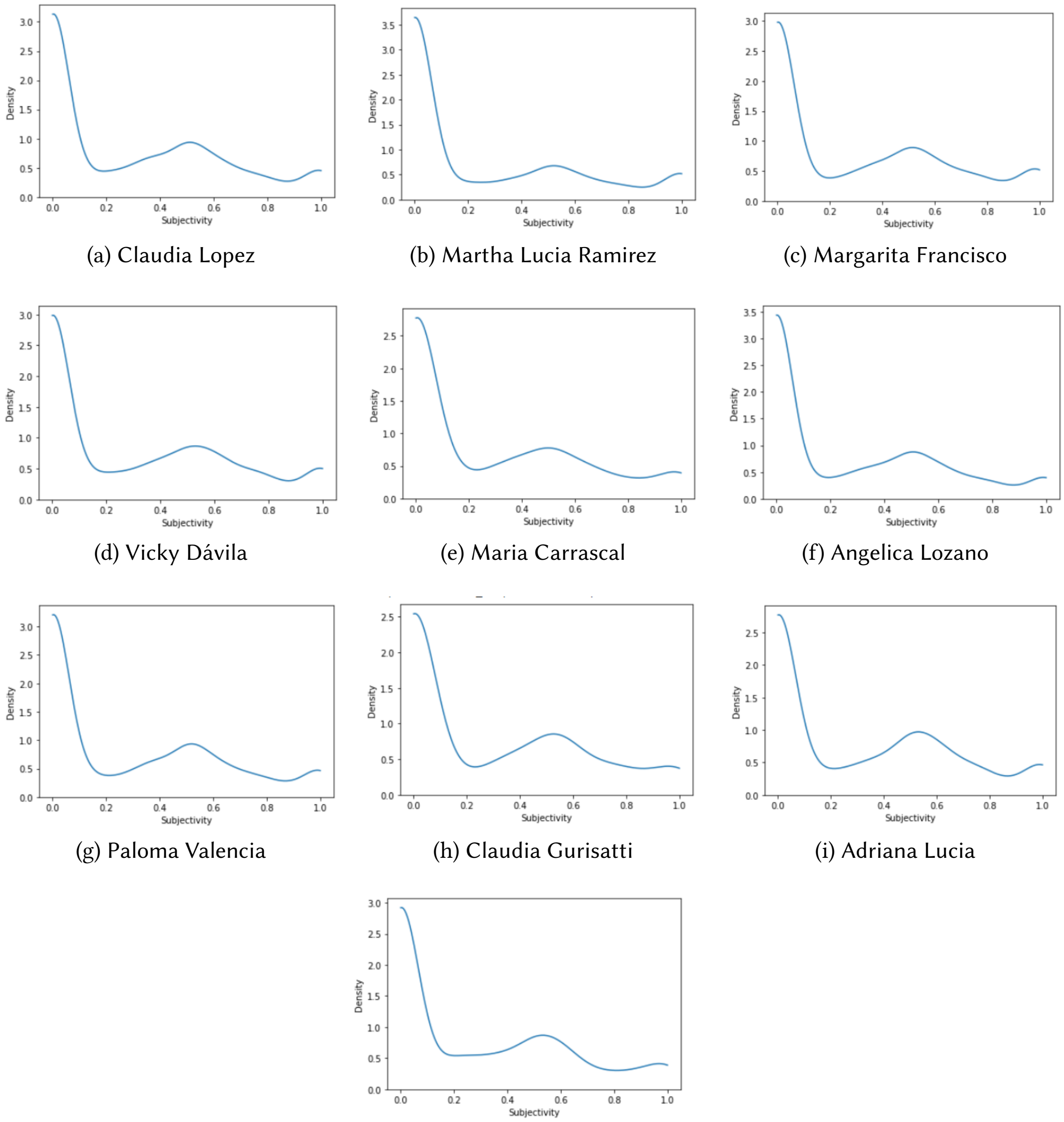
(g) Paloma Valencia

(h) Claudia Gurisatti

(i) Adriana Lucia

(j) Angela Robledo

Figure 4: Distribution Density of Polarity



(a) Claudia Lopez

(b) Martha Lucia Ramirez

(c) Margarita Francisco

(d) Vicky Dávila

(e) Maria Carrascal

(f) Angelica Lozano

(g) Paloma Valencia

(h) Claudia Gurisatti

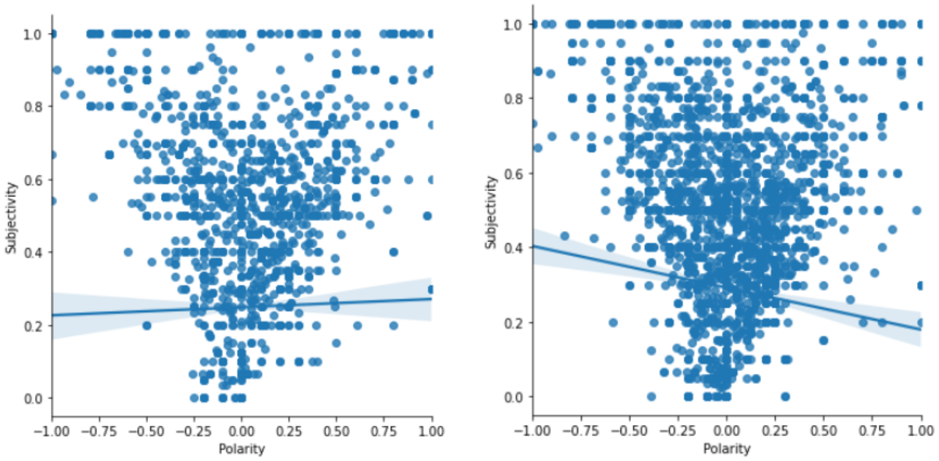
(i) Adriana Lucia

(j) Angela Robledo

Figure 5: Distribution Density of subjectivity

or personal judgement. In all cases, tweets tend to contain messages with factual information rather than subjective opinions.

Figure 6 presents shows the correlation between polarity and subjectivity. We observe that for Martha Lucia Ramirez, there is a no significant correlation between polarity and subjectivity. However, in the case of Paloma Valencia, there is a slight negative correlation between polarity and subjectivity, i.e., the more positive a tweet is, the more factual it tends to be.



(a) Martha Lucia Ramirez

(b) Paloma Valencia

Figure 6: Polarity vs. Subjetivity

7. Discussion

With the findings obtained, the presence of violent elements directed at female public figures in Colombia was notable. Taking into account the profiles of the public figures, it was evident that in the political sphere there is evidence of controversy by opinion. In summary, the texts aimed at the political profile tend to present a negative sentiment compared to the artistic profile. As mentioned by the Colombian Human Rights Ombudsman's Office in its 2021 annual bulletin, recently situations of psychological violence have worsened and concentrated in the Colombian population [3]. For this reason, it is considered that in the digital scene there is the presence of shocks of political thought leading to the formation of comments with a negative meaning directed at women of a public nature. This situation is a growing social problem, which gives way to the participation of technology as a basis for analysing the information generated on social networks such as Twitter, Facebook or Instagram.

It is evident that the coherence of a dialogue with good values must be satisfied with the absence of violence and the presence of respect. However, in this study the presence of negative elements towards women was observed, which is interpreted as a violation of human rights, autonomy and freedom of expression. With the existence of the Human Rights Ombudsman's Office as a control body in Colombia, an entity that currently has the function of issuing early warnings of violence in the field of the armed conflict and also in charge of preventing the violation of human rights. For this reason, the proposal of this article as a methodology for the monitoring of violence in the digital scenario, a potentially valuable source of information can be consolidated as support for the emission of early warnings of gender violence in social networks, with established criteria such as the frequency, intensity and quantity of negative opinions. After the deployment of this tool, it is hoped that it will be possible to contribute to the mitigation of the the impact of acts of psychological violence for social reconstruction in the Colombian population.

8. Conclusions

In this work, we identify elements of psychological violence against women in Colombia based on Twitter responses to public figures. The result of this study allows progress in social construction by providing tools to control institutions such as the Human Rights Ombudsman in Colombia for the prevention of potential human rights violations. We present a methodology to standardise the tweet extraction process, consolidate and analyse the information from the social network. This constitutes an instrument for the real dimensioning of psychological violence. The analysis carried out in said methodology includes the subjective component of a text, used for interpretation in relation to the positive or negative polarity found. Thus, with the adoption of the proposed methodology, the state control bodies and human rights organisations in Colombia and elsewhere can agree on a unified criterion of comparison, a fact that will contribute to the homogenisation of the protection, promulgation and prevention of human rights.

It is hoped that after applying the proposed methods for the analysis and having studied the data processed by the architecture proposed, it will be possible to work and involve experts in the field of women's rights and gender, experts in linguistics to optimise the tool and generate new strategies that promote human rights. Further work includes incorporating data from other other data sources such as Instagram, Facebook or WhatsApp, bearing in mind that to optimise data analysis it is appropriate to integrate data from diverse sources, since the problem studied is evident on the other social networks as well. We also consider that it would be fruitful to include a male control group, so as to eliminate references of non-validated sexist discrimination based on gender-neutral derogatory comments. Furthermore, we envisage that it would be appropriate to incorporate linguistic techniques that contribute to the more precise detection of misogyny in these digital settings, in addition to correlating other properties resulting from the integration with specialised areas in human rights and linguistics.

References

[1] U. Women, The world for women and girls annual report 2019-2020, 2020.

[2] M. I. L. Vélez, L. M. E. Jaramillo, Derechos laborales y de la seguridad social para las mujeres en colombia en cumplimiento de la ley 1257 de 2008, Revista de Derecho (2015) 269–296.

[3] D. D. Pueblo, Situación de las mujeres y personas con orientación sexual e identidad de género diversas, refugiadas y migrantes en colombia, Women's rights (2021) 10. URL: https://www.defensoria.gov.co/public/pdf/Boletin_Situacion_Mujer_2020.pdf.

[4] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, S. Belfkih, Big data technologies: A survey, Journal of King Saud University-Computer and Information Sciences 30 (2018) 431–448.

[5] S. Kemp, 2020, Digital 2020: 3.8 billion people use social media, URL: https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media.

[6] E. Van der Klashorst, S. Safarikova, Twitter as barometer of public opinion on the female athlete: The case of caster semenya, African Journal for Physical Activity and Health Sciences (AJPHES) 24 (2018) 649–658.

[7] A. Khatua, E. Cambria, A. Khatua, Sounds of silence breakers: Exploring sexual violence on twitter, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 397–400.

[8] S. D. M. Dussan, M. Leon, O. Garcia-Bedoya, I. Galpin, Exploring the colombian digital divide using moodle logs through supervised learning, Interactive Technology and Smart Education (2021).

[9] D. F. Martinez, J. N. Pacheco, L. F. Payan, F. C. Cepeda, Exploring the digital gender divide: Insights from the colombian case, IDIA2020 (2020) 69.

[10] Y. M. Shum, 2020, Situación digital, internet y redes sociales colombia 2020, URL: https://yiminshum.com/social-media-colombia-2020/.

[11] R. Wirth, J. Hipp, Crisp-dm: Towards a standard process model for data mining, in: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, volume 1, Springer-Verlag London, UK, 2000, pp. 1–11.

[12] S. Z. Cremades, Redes sociales para la prevención del suicidio juvenil, 3C TIC. Cuadernos de desarrollo aplicados a las TIC (2019) 54–69.

[13] J. R. Saura, A. Reyes-Menéndez, P. Palos-Sanchez, Un análisis de sentimiento en twitter con machine learning: Identificando el sentimiento sobre las ofertas de# blackfriday, Revista Espacios 39 (2018).

[14] J. C. Barriga Mariño, et al., Desarrollo y aplicación de una herramienta de extracción y almacenamiento de datos de twitter a un contexto social de violencia política, technology (2017).

[15] R. Silva, et al., Detección de violencia verbal hacia las mujeres en redes sociales mediante técnicas de aprendizaje automático, technology (2019).

[16] G. Evolvi, Hate in a tweet: Exploring internet-based islamophobic discourses, Religions 9 (2018) 307.

[17] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360.

[18] Google, 2021, Cloud firestore data model, URL: https://firebase.google.com/docs/firestore/data-model.

[19] G. A. García Vélez, Aplicación de la metodología crisp-dm a la recolección y análisis de datos georreferenciados desde twitter, technology (2018).

[20] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.

[21] F. I. Nicolai Manaut, Sistema de análisis de tópicos para interacciones cliente-call center, technology (2019).