Sentiment Analysis of Social Network Content to Characterize the Perception of Security

Luisa Fernanda Chaparro
Department of Mathematics
Universidad Nacional de Colombia
Bogota, Colombia
luchaparros@unal.edu.co

Cristian Pulido
Department of Mathematics
Universidad Nacional de Colombia
Bogota, Colombia
cpulido@unal.edu.co

Jorge Rudas
Institute of Biotechnology
Universidad Nacional de Colombia
Bogota, Colombia
jerudass@unal.edu.co

Ana Maria Reyes
Department of Mathematics
Universidad Nacional de Colombia
Bogota, Colombia
amreyesp@unal.edu.co

Jorge Victorino
Department of Computer Science
Universidad Nacional de Colombia
Bogota, Colombia
jevictorinog@unal.edu.co

Luz Ángela Narváez
Department of Mathematics
Universidad Nacional de Colombia
Bogota, Colombia
lanarvaezn@unal.edu.co

Francisco Gómez
Department of Mathematics
Universidad Nacional de Colombia
Bogota, Colombia
fagomezj@unal.edu.co

Darwin Martinez
Department of Mathematics
Universidad Nacional de Colombia
Bogota, Colombia
demartinezr@unal.edu.co

***Abstract*—The study of the perception of security helps to understand the feeling of citizens in the face of risk and the magnitude of its consequences, to understand the fear of crime. Traditionally the measure of this kind of subjective perception was made through surveys with low costly-effective performance, to a small sample of the population in specific time periods. To solve those inconveniences, this work analyzes take advantage of the amount of data and the real-time monitoring allowed by the social networks to quantifies the Perception of the security in Bogota. The quantification is made through different methods that involve ruled-based and supervised learning approaches for the sentiment analysis of the data coming from Spanish text from Twitter.**

***Index Terms*—Sentiment Analysis, Perception of Security (PoS), Natural Language Processing.**

I. Introduction

The study of the perception of security helps to understand the feeling of citizens in the face of risk and the magnitude of its consequences, to understand the fear of crime [1]. It means the emotional response (subjective perception) that an individual would face if he/she was a victim of crime. The fear of crime could change in time and depends on individual circumstances and experiences.

The perception of security turns out to be a difficult subject to quantify. Surveys have traditionally been used, which consolidate information on specific segments of the population at a particular period of time. The surveys are costly in resources and time and do not allow constant monitoring. A source of information that could help to solve these problems are the social networks.

Due to the advancement of technology, new digital platforms, and the fast growth of social networks, especially during the last two decades, the social behavior of the communities has acquired more importance. Nowadays, people connect with others in unimaginable ways, as individuals or collectives, allowing making friendships, belonging to interest topic groups, creating media content, and even influencing others' decisions and opinions in real-time [2], [3]. Social networks have the particularity to allow transmit events and news related to any field in real-time. The content generation and easy way to spread it through them, make the social networks the ideal source for obtaining large volumes of data that help to understand people's behavior. The work focuses on Twitter content, which not only has the advantages to be microblogging but also, to have a big amount of data due that counts with more than 150 million active users daily [4].

Twitter contents could explain the kind of response that a community has when a particular event surrounds them, how the people receive, process, and interpret the information coming from the media content. Tweets are valuable assets to understand people's perceptions [5], in particular, to characterize the citizen's perception of security. Nevertheless, the analysis coming from social networks and the impact that the contents have on spreading it are highly related to the tone in which the tweets were generated. For instance, the

Supported by Bank of National Investment Programs and Projects, National Planning Department, Government of Colombia (BPIN: 2016000100036)

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

people's perception and the feelings that could have in terms of the city's environment may result in the polarization of this perception [6]. Therefore, it is necessary to analyze the content's tone, which is called by experts as sentiment analysis [7]. Sentiment analysis has previously been used by different fields to know in real-time, for instance, the satisfaction of their customers [8], the impact their comments have on the image of their brands [9] even in the effect of the stock prices returns [10]. The sentiment analysis has been used for crime analysis, for instance, narcotics, criminal damage, burglary, hacking, among others [11], [12]. However, most of these works have focused on English. In the case of Spanish, until now the literature does not reflect works that require the analysis of the sentiment of the contents of social networks that quantifies the perception of the security of a city [13]. Because there are grammatically similar words between languages, the meaning of them and the colloquial use of some terms referring to a geographic location, in the particular case of study Bogota city, can provide us with an approach to correct indices in the perception of security.

This paper explores the performance of several methods and classifiers for sentiment analysis, categorizing the tone of security-related Bogotá related tweet content. Two methods are studied, a ruled-based and supervised method. The ruled-based is the traditional approach to the problem, using lexicons [12]. This approach is widely applied to English media content and is well documented in the literature. Nevertheless, in Spanish media content, the use of lexicons is limited due to the existing few of them. On the other hand, for the supervised method, machine learning techniques are implemented [13]. Understanding the operation and its performance, allows us to choose the best method to optimize the classification of the tweets based on their sentiment. This analysis allows quantifying in an automatized way the perception of security in the Bogota city case and monitoring in real-time.

II. Materials and Methods

A schematic representation of the method for identifying the sentiment of Tweets georeferenced in Bogota is shown in Fig.1. After the acquisition and preprocessing of the database, it's analyzed under two different approaches: ruled-based and supervised learning. For the first case, the classification is produced by comparison with defined word dictionaries. In the second case, the data is divided into training and testing sets, and the data is classified using various machine learning classifiers methods. For both cases, as a result, a score is obtained as a class. Finally, a performance analysis is done over the classification results.

A. Data Acquisition

The analyzed social network content comes from the social network Twitter. The test database contains around 26000 unique tweets, geo-located in the city of Bogotá, previously filtered, as a result of a media listening exercise carried out by the District Security Secretary from March 2019 until March 2020. To filter the tweets, the Security Secretary defined a set of words related with security issues, into that set, the words that can find are "robo, asalto, inseguridad, seguridad, arma, atraco, rateros, alarma, violación", among others, are words, whose meaning is related to security issues and allow differentiation between the real and the noise content. The final database contains tweets that at least have one or more of those filter words. To have a comparison basis, the whole database was additionally labeled according to the sentiment of each tweet. To have a comparison basis, the database was additionally labeled according to the sentiment of each tweet, by a group of experts. The experts scored each tweet, according to the sentiment in was originally written. They are around 10 qualified people that work with the Security Secretary and with the analytics team at Universidad Nacional de Colombia.

B. Pre-processing

For preprocessing and due to the nature of the tweets, links, mentions, and hashtags were removed [14]. Likewise, the texts were normalized by standardizing capitalization and eliminating punctuation marks. It should be remembered that since it is an analysis of texts in Spanish, both final and initial punctuation marks are considered, the accents of the language are eliminated and stopwords are discarded [15]. As the meaning of the words is contained in the root the words are stemmed. At the end of the preprocessing, there is an irrelevant wordless database, suitable for tokenization, vectorization, and subsequent sentiment analysis through the two main methodologies: rule-based and machine learning.

C. Rule-based Method: Lexicon

Rule-based sentiment analysis is defined as the study conducted by the language experts for determining the tone and polarity into the texts. As a result, a set of rules (also known as a lexicon) according to which the words classified are either positive or negative along with their corresponding intensity score is obtained [16].

For the Ruled-base method, two-word lists (positive and negative) are defined with the help of a lexicon, which contains a previous classification. These lists are known as dictionaries. The lexicon used for these experiments was *ML-Senticon*, developed by the University of Seville [17]. The tweet content is compared with the list of words provided by the dictionaries that are previously defined. Depending on the way to obtain the score provided by the dictionaries, is assigned a number for each word and the total score is added for each tweet.

There are two ways to obtain the score: a simple score and a polarity score. For the simple score, only is taken into account the appearance frequency of the word. If the word is defined in the dictionaries as negative the score sum -1 and 1 for the positive case. In that case, the score sums the number of positive words ($Pos\ w$) and the negative ones ($Neg\ w$), as is shown:

$$Simple\ Score_{(tweet)} = \frac{\sum_{w \in words} Pos\ w - Neg\ w}{\sum w} \quad (1)$$

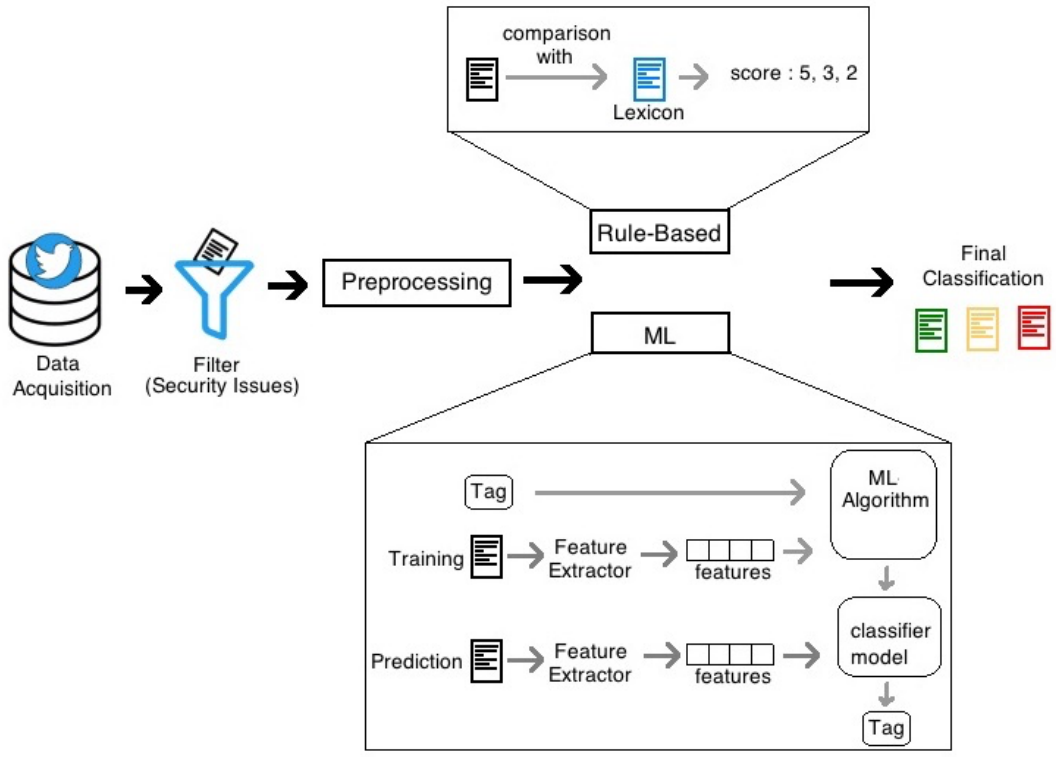

Fig. 1. Schematic Representation of the Sentiment Analysis over Tweets. Start with data acquisition followed by a filter of relevant data. The data pass over a preprocessing step to implement the different sentiment Analysis approaches. As a result, a final classification: positive, negative, or neutral for each Tweet.

On the other hand, for the polarity score (Equation 2), the score given by the dictionary for each word is considered, as follows:

$$Polarity\ Score_{(tweet)} = \frac{\sum_{w \in words} S\ Pos\ w + S\ Neg\ w}{\sum w} \tag{2}$$

where S is the score of the polarity in the dictionaries.

To obtain the tweet score, the score of individual words is summed. It means that a tweet that is counting as a positive tweet with a simple score, could be more positive or more negative depending on the meaning of the words used on it, if we use the second method. This is because for the second case the score is based on the lemma of the words. For each tweet, the final sentiment will be given by the total score, divided by the total number of words within the tweet.

D. Supervised Learning Methods

1) Features Extraction: For the Supervised machine learning methods, two models were used to determine the dictionaries: Bag Of Words and TF-IDF. A Bag of Words (BOW) is a representation of text that describes the occurrence of words within a document. In this model, a text like a tweet is represented as the bag of its words. This approach does not take into account the grammar or word order but includes the multiplicity [18]. This model is commonly used in methods of document classification, where the frequency of occurrence of each word is used as a feature for training a classifier like Multinomial Naive Bayes or a Logistic Regression [19]. On the other hand, the TF-IDF model (Term Frequency - Inverse Document Frequency) is a method for extracting characteristics in a corpus, which it provides is a numerical measure that expresses how relevant a word is to a document in a collection. The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the document collection, allowing you to handle the fact that some words are more common than others [20]. TF-IDF is the product of two measurements, term frequency, and document inverse frequency. The first is a score for the frequency of the word in the current document, while the second is a score for how rare the word is in the documents, or better, how important it is.

The term frequency can be determined as:

$$tf_{t,d} = \frac{n_{(t,d)}}{number\ of\ terms\ in\ document} \tag{3}$$

Where n is the number of times the term t appears in document d. The inverse frequency IDF can calculate:

$$idf_t = \log\left(\frac{\#\ of\ documents}{\#\ of\ documents\ with\ the\ term\ t}\right) \tag{4}$$

2) Classifier Methods: In supervised classification methods, labeled documents are grouped into predetermined classes. It means that a model can be constructed according to existing samples based on which unlabeled data assigned

to their respective categories [21]. Then get the methods from the dictionaries, proceed to test, and test various classifiers, to compare the performance between them and with the rule-based method. For these experiments, several classification strategies were explored, including, Multinomial Naive Bayes (MNB), Logistic Regression (LR), Bernoulli Naive Bayes (BNB) and Stochastic Gradient Descent (SGD) Classifier.

3) Validation: For the validation of the Machine learning classifiers a Cross-validation was implemented. This type of validation is a statistical method used to estimate the skill of machine learning models and their consistency [22].

The algorithm uses a single parameter called k that refers to the number of sets into which the data will be divided. This is why it is called k times cross validation. The validation implemented in this work performs the validation for 50 times, that is a k = 50. This type of cross-validation is used in applied machine learning to estimate the ability of a machine learning model on invisible data. In other words, use a limited sample to estimate how the model is expected to perform in general when used to make predictions on tests and not training data.

The general procedure of the algorithm is as follows [23]: First, the date is mixed randomly. The data set is divided into k groups (50 in this case). For each unique group, the group is taken as a test data set, the other groups will be taken as training. After the division of the dataset, the model is trained and evaluated. To measure the performance of each of these models a group of metrics is obtained. These metrics are accuracy, precision, f1, and recall and are related to the confusion metrics (false negatives, false positive, true negative, and true positive).

Accuracy is the most intuitive performance measure. Its a ratio between the correctly predicted observation over the total. This metric works well if the number of false-positive and false negatives is almost the same in the results. Precision gives the ratio of correctly predicted positive observations over the total predicted positive. If the database shows high precision, means that there is a low false positive. To measure the sensitivity of the results, a Recall metric is calculated. This measure gives the idea of the ratio of correctly predicted positive into all observations of a determined class. Finally, the F1 score is the weighted average between precision and recall [23].

4) Statistical Analysis: A statistical significance analysis was performed over the metrics obtained previously. For each metric, a p-value of 0.05 and 0.005 is calculated. The p-value gives the probability of the results of the metrics deviating.

III. Results

All the 26256 tweets acquired talking about security issues, nevertheless, the tone of writing is so different between them making the quantification of the perception of security a challenging job. Table I, shows a sample of tweets acquire during the listening media exercise.

After the initial data collection and filtering, it is important to have a gold standard as a comparison parameter for the sentiment analysis models that are implemented. This Gold-Standard is produced by a manual tagging, tweet by tweet, by the group of experts. The distribution of the score made by the experts in the whole database are shown in Fig. 2.

#Horrible es movilizarse en Bogotá y encontrar todas las semanas una marcha. Eso si es #Horrible tener que aguantar a los desadaptados.
#BogotaInternacional || También conocieron los avances en movilidad de la ciudad, a través de las diferentes estrategias implementadas por @SectorMovilidad para mejorar la movilidad en proyectos en actores claves como: peatones, bicicletas y seguridad vial.
En @SectorMovilidad trabajamos para que las mujeres en Bogota puedan disfrutar de una movilidad limpia, segura y eficiente. El secretario de Movilidad,@nico_estupinan, explica.
Horror en Bogotá: otro caso de intento de secuestro de un menor causó pánico en la capital.
Deja que entre el ESMAD, la policía, yo que sé, pero eso de que estos criminales usen la universidad como escudo tiene que acabarse.La autonomía universitaria que tanto invocan se violó hace raaaattttto desde que esta gentuza hace lo que quiere con impunidad.
Envían a prisión a ocho personas dedicadas al robo de celulares en Transmilenio
Se fortalece la capacitación de los nuevos #AuxiliaresDePolicíaMEBOG a través del Programa Mejor Policía de @Bogota @SeguridadBOG. #ConstruyendoSeguridad
Trabajando en equipo con @PoliciaBogota en los puntos de prevención y control de los delitos en @Bogota . Seguimos consolidando política basada en evidencia. @SeguridadBOG
Es triste llegar al punto en el que tus seres queridos te digan: "por favor no se exponga más en bici, lo queremos con vida". @Bogota @PoliciaBogota necesitamos acciones reales para frenar el hurto y asesinato a bici usuarios.

TABLE I
A SAMPLE OF TWEETS ACQUIRE DURING THE MEDIA LISTENING EXERCISE CARRIED OUT BY THE DISTRICT SECURITY SECRETARY OF BOGOTA BETWEEN MARCH 2019 AND MARCH 2020.

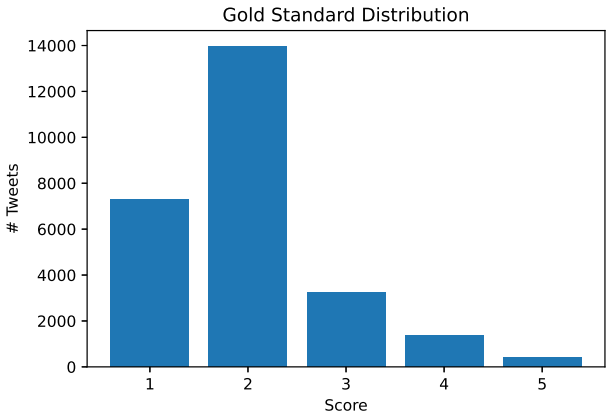

Fig. 2. Distribution of the score made by the experts over the database. The experts scoring each tweet, according to the sentiment in were originally written on a scale from 1 to 5, where 1 means a negative sentiment (like anger or frustration) and 5 a positive sentiment (like optimism or happiness).

At the end of the preprocessing step, and before the stemmed process, each tweet is divided word by word, this is known as tokenization. A descriptive visualization of these words can be seen in Fig. 3. There is a frequency distribution of the 100 most frequent tokens in the database.

For both kinds of feature extract sentiment analysis methods, Ruled-based and supervised machine learning, a set of metrics is obtained.

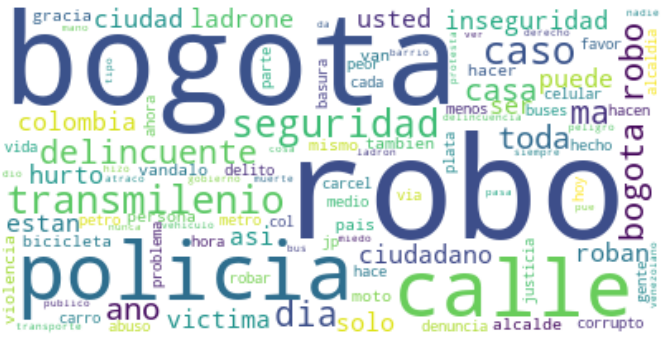

Fig. 3. The 80 most frequent tokens in the database as a word cloud. The size of the word represents a higher frequency in the database. Those tokens show that the tweets not only speak about security issues but also about localization and places into the city.

For the rule-based method, a unique value for the metrics is obtained in each of the ways to get the score. The performance measures are shown in Fig.4. As expected for the Polarity

Fig. 4. Metrics for the Rule-based method: Accuracy, Precision, F1 Score and Recall are shown for a polarity Score calculation (Blue) and Simple Score Calculation (Red).

score case, the individual metrics (accuracy and precision) show a better performance than for the simple score. It is due to the particular score get for each word due to the meaning of them, in other words, due to the lemma of the words and in this way, the method classifies nearly the same as a human. Nevertheless, the Polarity Score in combination shows a worse performance than the Simple Score, this behavior is due that the lexicon does not take into account the meaning of the neighbor's words.

For the supervised method, after the K-Fold cross-validation, we obtained for each model and classifier a set of metrics. Nevertheless, to understand and choose the classifier with the best performance, we use the values for the first quartile. These metrics are shown in Fig.5.

In the case of Bernoulli Naive Bayes classifier, the behavior is indifferent to the vectorization method with one of the lowest accuracy. The accuracy is quite similar for all the classifiers in both vectorizing methods, nevertheless, the Logistic Regression and the Multinomial Naive Bayes give the best rates on precision. The Logistic Regression has a lower chance to deviate according to the p-value calculated. Both classifiers also give an idea of low misclassification rates, especially in the case of Multinomial Naive Bayes that also shows a smaller chance of deviation of the results.

IV. Discussion

Social media is rich in both content and relationships, leading to challenges and great opportunities in sentiment analysis. Although the relationships of social networks have been widely discussed from social points of view (psychological and sociological), their understanding from a machine learning approach is in its early stages, especially if we talk about content in Spanish.

This first approximation to the classification of feelings from the tweets generated in the city of Bogota, allows this analysis to quantify the perception of citizens in a better way than as it was traditionally done through surveys. This is the first time that the perception is analyzed in Bogota using social media content. It allows analyze a greater amount of information and make this exercise continuous over time. Although there is literature that talks about sentiment analysis, this is the first work that focuses on the analysis of sentiments on security issues in a city and the perception of fear of crime.

The classification of texts and the feelings that come from them require supervised learning, where the orientation of feelings must be known a priori to obtain specific predictive models. However, we have studied the differences and the information we can obtain in rule-based and supervised methods so that we can automate the process in the best way and obtain the best performance as if it were done by groups of experts.

The supervised methods have a better performance in the classification of the feelings of the texts, over those based on rules, because they not only take into account the content of the texts but the relationship that the words have in them. Regarding rule-based models, they present challenges in the design of lexicons, which depend on depth layers in terms and language. Lexicons in Spanish are poorly developed, which can lead to lower performance in the Sentiment analysis, as we can see when comparing the results of the metrics. In addition, the specificity in the corpus terms would be improved as soon as the database increases, and some events that at the moment are not listed will be on future tweets.

Comparing the supervised methods, so far, the Multinomial Naive Bayes using the TFIDF vectorization, gives the best precision classification performance, allowing have a first attempt to quantify the perception of the security in Bogota. Nevertheless, the f1 score that involves not only the precision but also the sensitivity shows that any other classifiers show a better performance than the Multinomial Naive Bayes-TFIDF. That results give us two approaches into consideration: one which takes into account precision as a performance measure and one in which also considers the sensitivity.

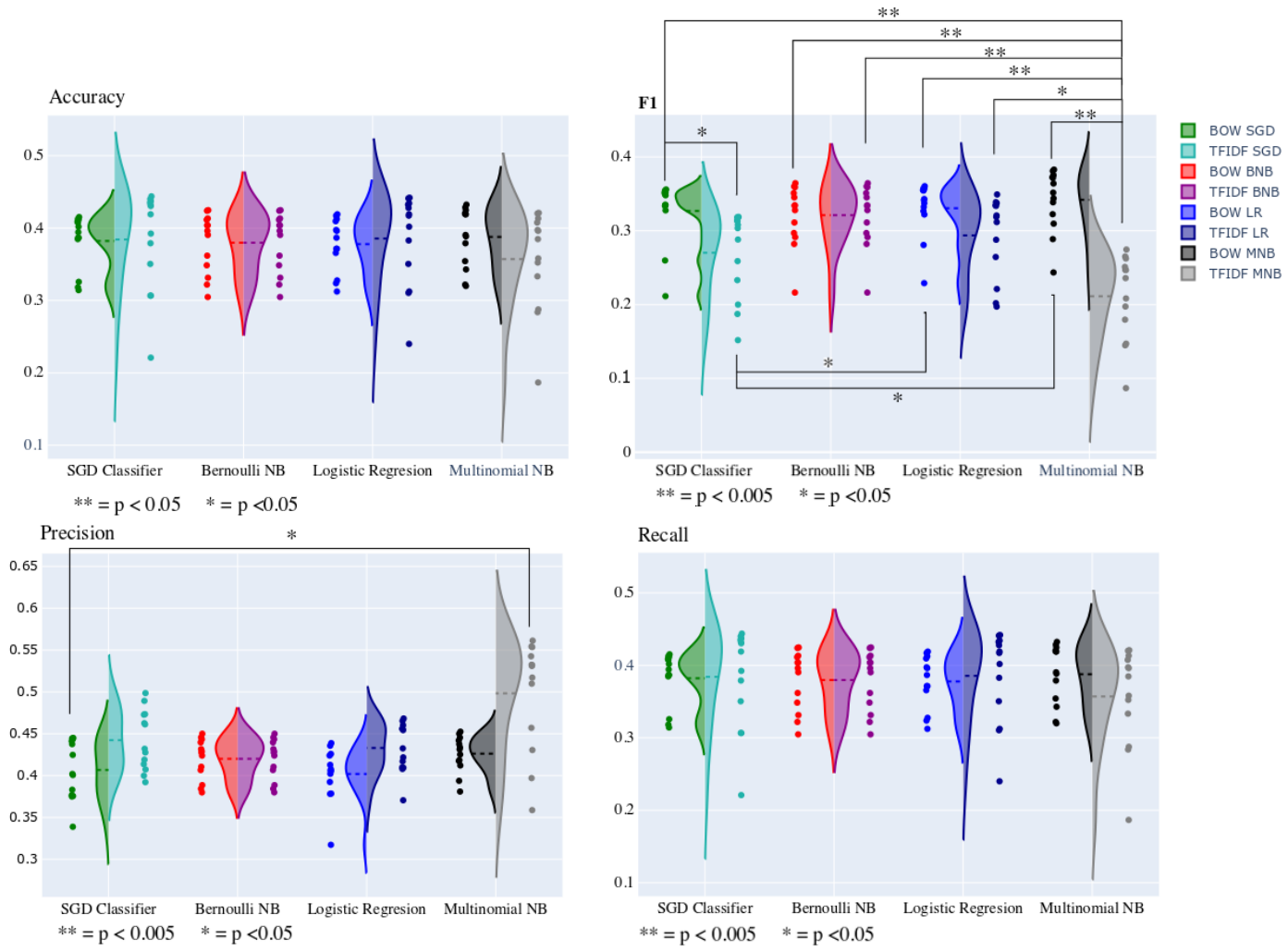

Fig. 5. Metric for all the classifiers (SGD: Stochastic Gradient Descent, BNB: Bernoulli Naive Bayes, LR: Logistic Regression, MNB: Multinomial Naive Bayes) in both vectorizing models: BOW (Bag of Words) and TF-IDF. (Upper Left) Accuracy, (Upper Right) F1 Score, (Down Left) Precision, (Down Right) Recall. All the metrics are plotted for the first quartile of data. In all the cases p-value of 0.05 and 0.005 are calculated.

These results will be used as an input to a model in which the numbers of followers and the retweet rate and like will be considered as a complementary measure of the PoS. In the case in which only the model takes into account the precision, the Multinomial Naive Bayes-TFIDF is used as a measure of the PoS. Nevertheless, taking into account the second approach and care also the sensitivity, the better performance is shown by the SGD Classifiers using again TFIDF, which shows a good F1 and also the second-best precision performance will be used. To understand the criteria of selection behind the classifier, the results must be analyzed by an interpretability model, such as LIME (Local Interpretability Model Explanation).

Acknowledgements

This work was funded by the project "Diseño y validación de modelos de analítica predictiva de fenómenos de seguridad y convivencia para la toma de decisiones en Bogotá", at Bank of National Investment Programs and Projects, National Planning Department, Government of Colombia (BPIN: 2016000100036).

References

[1] T. Rundmo and B. Moen, "Risk perception and demand for risk mitigation in transport: A comparison of lay people, politicians and experts," *Journal of Risk Research*, vol. 9, pp. 623–640, 09 2006.

[2] M. Arroyo Lazo, "Schwab, klaus. the fourth industrial revolution. ginebra: World economic forum, 2016, 172 pp.," *Economia*, vol. 41, pp. 194–197, Oct. 2018.

[3] Schultz-Jones Barbara, "Examining information behavior through social networks: An interdisciplinary review," *Journal of Documentation*, vol. 65, pp. 592–631, Jan. 2009. Publisher: Emerald Group Publishing Limited.

[4] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: An analysis of a microblogging community," in *Advances in Web Mining and Web Usage Analysis* (H. Zhang, M. Spiliopoulou, B. Mobasher, C. L. Giles, A. McCallum, O. Nasraoui, J. Srivastava, and J. Yen, eds.), (Berlin, Heidelberg), pp. 118–138, Springer Berlin Heidelberg, 2009.

[5] M. E. Brown, P. A. Dustman, and J. J. Barthelemy, "Twitter impact on a community trauma: An examination of who, what, and why it radiated," *Journal of community psychology*, February 2020.

[6] A. Alamsyah and F. Adityawarman, "Hybrid sentiment and network analysis of social opinion polarization," in *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, pp. 1–6, 2017.
[7] B. Liu, *Opinions, Sentiment, and Emotion in Text*. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015.
[8] J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *JASIST*, vol. 60, pp. 2169–2188, 11 2009.
[9] K. Curran, K. O'Hara, and S. O'Brien, "The role of twitter in the world of business," *IJBDCN*, vol. 7, pp. 1–15, 07 2011.
[10] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič, "The effects of twitter sentiment on stock price returns," *PloS one*, vol. 10, p. e0138441, 09 2015.
[11] X. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using twitter sentiment and weather," in *2015 Systems and Information Engineering Design Symposium*, pp. 63–68, 2015.
[12] B. R. Prathap and K. Ramesha, "Twitter sentiment for analysing different types of crimes," in *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pp. 483–488, 2018.
[13] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *Social Computing, Behavioral - Cultural Modeling and Prediction* (S. J. Yang, A. M. Greenberg, and M. Endsley, eds.), (Berlin, Heidelberg), pp. 231–238, Springer Berlin Heidelberg, 2012.
[14] N. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, pp. 170 – 179, 2014.
[15] S. Bhagvat, ""clustering of twitter technology tweets and the impact of stopwords on clusters"," Master's thesis, San José State University, 2011.
[16] V. Singh, G. Singh, P. Rastogi, and D. Deswal, "Sentiment analysis using lexicon based approach," in *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp. 13–18, 2018.
[17] F. Cruz, J. Troyano, B. Pontes, and F. J. Ortega, "Ml-senticon: A multilingual, lemma-level sentiment lexicon," *Procesamiento de Lenguaje Natural*, vol. 53, pp. 113–120, 09 2014.
[18] S. K and S. Joseph, "Text classification by augmenting bag of words (bow) representation with co-occurrence feature," *IOSR Journal of Computer Engineering*, vol. 16, pp. 34–38, 01 2014.
[19] A. Basarkar, ""document classification using machine learning"," Master's thesis, San José State University, 2017.
[20] S. Qaiser and R. Ali, "Text mining: Use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, 07 2018.
[21] T. Ayodele, *Types of Machine Learning Algorithms*. 02 2010.
[22] P. Refaeilzadeh, L. Tang, and H. Liu, *Cross-Validation*, pp. 532–538. Boston, MA: Springer US, 2009.
[23] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," vol. 3408, pp. 345–359, 04 2005.