On the power of social networks to analyze threatening trends

Julián Ramírez Sánchez, Alejandra Campo-Archbold, Andrés Zapata Rozo, Daniel Díaz-López
School of Engineering, Science and Technology, Universidad del Rosario, Bogotá, Colombia
{julians.ramirez, alejandra.campo, andresf.zapata, danielo.diaz}@urosario.edu.co

Javier Pastor-Galindo, Félix Gómez Mármol
Department of Information and Communications Engineering, University of Murcia, Spain
{javierpg, felixgm}@um.es

Julián Aponte Díaz
Armada Nacional de Colombia, Bogotá, Colombia
julian.aponte@armada.mil.co

***Abstract*—Despite their numerous advantages, social networks may be unfortunately employed for hate promotion or violence instigation. Thus, Law Enforcement Agencies (LEAs) must pay close attention to these platforms to preserve public safety and detect threatening trends. To this extent, the paper at hand discusses the role of social networks in today's social movements and presents the associated opportunities for monitoring, forecasting, and prevention. We document a practical use case, developed in collaboration with the Colombian Army, in which social media analysis and artificial intelligence play a crucial role in characterizing a national crisis.**

Index Terms: Social media, Cyberdefence, NLP, LEA, HSM.

1. The role of social media in civic movements

The recent transformation of social networks has been notable, evolving from interaction channels with people connecting individually with others to scenarios where a single individual can incite attitudes, thoughts, and ideas massively [1]. This influencing power is maximized by the capacities that social networks offer to connect with other people, e.g., the virtually unlimited capacity to fill the feeds of users from all over the world.

Some social networks may offer certain anonymity with which people can express ideas with a great sense of freedom without the fear of being judged [2]. However, people may feel social pressure to express their true thoughts when anonymity is not guaranteed [3].

Social networks may also be considered a parallel world to reality, where each person can find contacts with similar ideas regarding political, social, or economic positions [4]. In fact, there are specific functionalities to create communities of people with common interests, such as Facebook groups. A community may also be

observed in a group of users following the same Twitter or Instagram account [5].

Communities in online social networks may also impact the physical world and vice versa [6]. Messages or information promoted by certain communities in the cyberspace can finally end up in specific actions in the physical world [7]. One tangible example is the "*Black Lives Matter*" trend, which had its first appearance in 2013 [8]. This movement aimed to enforce the rights of the African-American community and used online social networks to promote protests in the streets of different countries of the world. Relation between physical world and cyberspace may be considered so strong that the absence of social media may even result in 10% fewer incidents in real life [3].

These representative events prove that digital platforms have become an influential ingredient for promoting civic unrest, empowering political protests, and keeping social movements alive. Online communication serves as a tool for coordinating protest activities, creating networks of support or opposition to ideological manifestations, and raising sensitization in society that may determine the success or failure of objections [6].

2. Links between digital and real world

Given the catalytic medium in the distribution of thoughts and mass mobilization that constitutes social media, it becomes ideal for investigating human behaviour and social phenomena [1]. Social media streams have been extensively analyzed for elections, market trends, public health, inference of user characteristics, and threat detection [9], [10]. Focusing on security, LEAs increasingly benefit from Internet-based investigation and Open Source Intelligence (OSINT) [2]. First, *online-to-online* operations could be defined as mining indicators in social platforms to identify signs of cybercrime, hate speech, harassment, or extremist within the virtual ecosystem [5]. The most common techniques include Natural Language Processing (NLP) and classifiers to spot malicious messages and identify perpetrators and at-risks users [11].

Second, we distinguish *offline-to-online* scenarios where protests or riots in the streets attract the attention of digital fora, provoking criticism, agitation, and polarization in social networks [3]. The mapping between the two worlds, the real and the digital, is commonly conducted through temporal correlations between physical events and themes extracted from social data streams through hashtag analysis, bag-of-words (BOW) techniques, or topical extraction algorithms [8].

Finally, *online-to-offline* activities would collect and inspect social media data to guarantee safety in real life [3]. Most commonly, we find studies that apply statistical analysis and data analytic to open data to understand the causes and consequences of an event or conflict [6]. Data scientists may early detect social movements with supervised or unsupervised learning, using common features such as texts, tags, locations, hours, or user metadata [4] to warn the competent authorities for a quick reaction in urban and rural terrains.

The above efforts should leverage event prediction models, highly differential for LEAs but also challenging for researchers [6]. In this regard, AI is indeed promising, with a community proposing several solutions based on regression, clustering, support vector machines, NLP, and network analysis, among other techniques [12]. In a complementary manner, social media metadata, semantic matching, and spatial-temporal analysis remain indispensable resources for these forecasting tasks [10].

3. Computing aggressive groups of protesters

This section presents a practical case in collaboration with the Colombian army that demonstrates the importance of social network analysis from a cyberintelligence perspective.

In the first semester of 2021, Colombia witnessed a social crisis characterized by protests in many Colombian cities. Some of them ended up in violent actions against public infrastructures such as police stations, government administration buildings, public transport stations, among others. In that crisis, social networks played a crucial role, collecting many of the opinions of the population and evidencing situations not shown in the principal national TV channels.

With the aim to analyze that social phenomenon, we applied the methodology presented in [13] that detects violent movements in Twitter using similarity models, sentiment analysis, clus-

tering, and graphs. We present the steps carried out in the online investigation in the following.

3.1. Collection of evidences

First, we gathered the data generated in social networks regarding such Colombian crisis. Specifically, we obtained the information available on Twitter as it is a social network with a potential audience of 3.35 million persons, a strong cultural influence, and a relevant use (59.2%) by people between 16 and 64 years in Colombia [14].

Thus, we used the `TAGS`[1] tool to collect tweets containing the following hashtags observed in the previous days of the protest: `#ParoNacional9J`, `#CIDHEscucheALasVíctimas`, `#CidhEscucheAlasVictimas`, `#SOSColombiaDDHH`, `#DuqueRenuncieYa`, `#DuqueDictador`, `#ColombiaEnAlertaRoja`, `#ColombiaSOS`, `#ColombiaResiste`, `#ColombiaSOSDDHH`, `#ColombiaSomosTodos`, `#ColombiaEnDictadura`, `#DuqueAsesino`, `#SOSColombiaNosEstanMatando`, `#PrimeraLineaJuridica`, `#PrimeraLinea`, `#CaliEnPeligro`, `#CaliSOS`, `#CaliResiste`, `#SOSCALICOLOMBIA` and `#ColombiaResiste`.

In total, we collected $15,034$ tweets from June 6th to June 11th, 2021, which include the previous and following days to the protest on June 9th, 2021. This sample of tweets was filtered not to contain retweets andpreprocessed by removing hashtags and mentions. Additionally, emojis were replaced by their meaning `emoji`[2] to avoid losing information. Finally, the Spanish tweets were translated to English using the `Google Translator`[3] python library to get a total of $1,601$ original and preprocessed tweets. A sample of them were manually verified to confirm its successful performance. Nevertheless, some particular expressions and idioms of Colombian Spanish are not solved by the translation service offered by Google Cloud. Thus, improvements in this regard would imply using a translator customized for Colombian Spanish that generates equivalent expressions in English.

3.2. Identification of related content

To avoid handling messages individually and to be able to make an aggregated analysis, it is interesting to calculate the similarity between tweets.

Thus, we vectorized the previously gathered tweets through the tool *word2vec*[4]. This processing produced a similarity matrix using the distance existing between the vector representation of each pair of tweets. The similarity matrix was composed of the similarity metric calculated using the cosine distance. In this way, the component in the position (i, j) of the matrix contains the distance between the tweets i and j. It is worth mentioning that *word2vec* uses an embedding from English words found in Google news, which differs from our context. A Colombian Spanish-based embedding would have a better performance, but currently, no such embedding shows good quality, volume, and veracity. Despite this, the applied vectorization still captures the primary meaning of each tweet.

3.3. Grouping of similar content

Once calculated the similarity of each tweet, we identified threads of messages around the same topics. Clustering is a technique where unsupervised machine learning models identify similar data groups. When clustering is applied to a similarity matrix, like the one obtained from the previous section, we obtain interesting groups of tweets that share a common meaning.

In our Colombian use case, we employed K-Means clustering to obtain groups of tweets with similar content. We used the Elbow Method and the Calinski-Harabasz index to determine the optimal number of clusters, experimenting from 1 cluster up to 8 clusters, to obtain an optimal value of $K = 3$ finally. The Elbow method allows to find heuristically the optimal number of clusters using two concepts: i) the distortion, which is the average of the squared distances between clusters

[1] https://tags.hawksey.info
[2] https://pypi.org/project/emoji
[3] https://cloud.google.com/translate
[4] https://code.google.com/archive/p/word2vec

centroids, and ii) the inertia, which measures the distance of each point from its nearest centroid. So, the correct number of clusters with the Elbow method is identified when distortion and inertia are low, and the ratio (distortion/inertia) drops linearly, i.e., $K = 3$, as shown in Figure 1. The latter is the point of balance (elbow) between the sum of squared errors (SSE) and the number of clusters in which both values are as low as possible.

On the other hand, the Calinski-Harabasz index is convenient in situations when the clusters are less spherical, such as in our case. The Calinski-Harabasz index is obtained from i) the ratio of the sum of squares in a cluster, and ii) the sum of squares of the distance between each cluster centroid. A higher ratio value means that the clusters are well separated and compact. In this regard, Figure 1 also shows a clear peak for the Calinski-Harabasz index when $K = 3$..

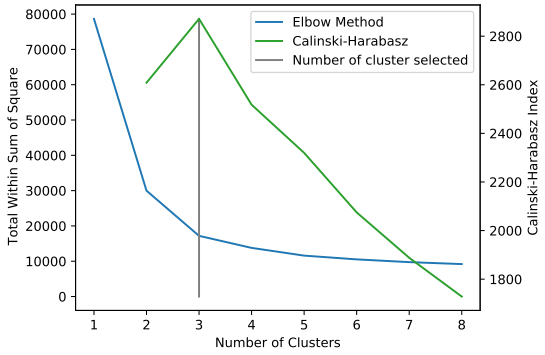

Figure 1. Number of clusters determined by the Elbow method and the Calinski-Harabasz index

To visualize these clusters in two dimensions, methods such as Principal Component Analysis (PCA) extracts the most significant features of data instances to be separated through clusters. In Figure 2, $pca1$ and $pca2$ are the first two principal components of obtained by a PCA for our dataset, representing the greatest variance of tweets in two dimensions and grouped in per cluster.

We better characterized the formed clusters in terms of the frequency of the words, so three different word maps were obtained, as reflected in Figure 2. The cluster with the highest number of tweets was cluster 0, which contains $1,020$ tweets mainly informative putting in evidence confrontations in popular districts in Cali and Bogotá. Follow cluster 2 with 431 elements, which represent mainly opinions with high subjectivity, some of them reacting in an aggressive way against actors of the conflict such as National Police, politicians and private organizations. Finally, cluster 1 has 150 tweets, containing mainly measured opinions and reactions.

3.4. Measurement of sentiment polarization

An interesting technique used to understand the meaning of phrases is the application of sentiment analysis. With these models, it is possible to detect the sentiment polarity depending on whether such a phrase evokes something positive, neutral, or negative.

Therefore, each cluster of tweets previously identified was analyzed in terms of sentiment polarity through the model of sentiment analysis provided by the Python library `TextBlob`[5]. Technically, the latter uses a pre-trained Single Layer Perceptron (SLP) classifier to determine a value of sentiment polarity $\rho \in [-1, 1]$. In this way, we consider a tweet as negative if $\rho < 0$, neutral if $\rho = 0$, and positive if $\rho > 0$.

The results of the sentiment analysis for each cluster are shown in Figure 3, which contains an example of a tweet for each cluster in Spanish and English versions. Particularly, the sentiment polarity for those exemplified tweets are -0.5 , 0 and -1 , respectively. Cluster 0 has the highest portion of negative tweets (23.3%), and cluster 2 has fewer but the most negative tweets, in the scale $[-1, 1]$. It is worth mentioning that negative phrases will not be necessarily violent, but violent phrases will generally be considered negative by a model of sentiment analysis.

3.5. Extraction of extreme communities

In this way, the cluster with the most negative tweets (cluster 2) was selected to be further reviewed in detail through the construction of a graph. Each user is represented by a node and the connections between nodes represent follower and mutual following relationships in Twitter depending on the existence of single or double arrows. Information regarding the relationships between the node in the graph was obtained using the `TinfoLeak`[6] intelligence tool.

[5]https://pypi.org/project/textblob/
[6]https://github.com/vaguileradiaz/tinfoleak

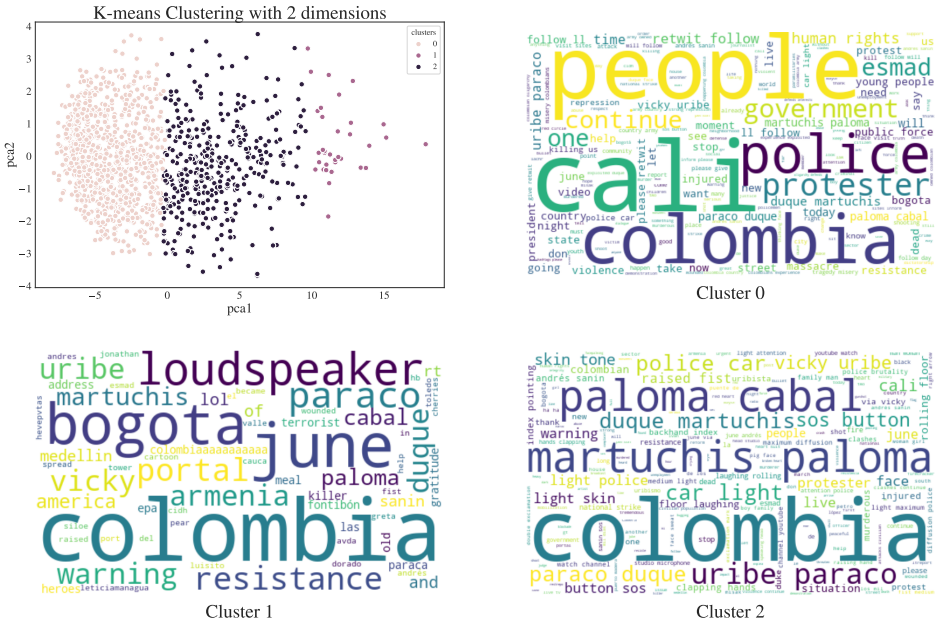

Figure 2. Three clusters of tweets with their corresponding word maps identified from the analysis of similarities

In this regard, the upper part of Figure 4 shows the graph for the Colombian use case. It was created using `Gephi`[7] after applying a k-core filter with parameter $k = 2$, which generates a subgraph where all vertices are connected and have a degree equal or greater than k. Such k value was selected because we are interested in users interacting with at least two Twitter accounts.

In the graph, we can observe two "isolated communities" where a node belonging to cluster 2 (a red one) is connected only to nodes that do not belong to any cluster (gray ones). We note that such segregated communities are either focused around protests in small towns or focused in violent incidents which did have national media impact. Additionally, we can also observe "connected communities" (main connections highlighted in blue) where nodes belonging to cluster 2 (red ones) are connected directly between them or through a third node. Such "connected communities" observed in the graph may be seen clearer in the lower part of Figure 4 and generally refers to discussions about incidents with high media impact such as murders of protestants, fires over main transport systems, etc.

The analysis of this graph allows us to pinpoint prominent and influencer nodes, with thousand of followers, which distribute information related to the protest. We can even spot users that had not been considered before because no tweet from them was collected (the gray nodes) but have a significant influence. For example, the node with an orange edge may be considered suspicious as it has around 18,000 followers and is related to several nodes from cluster 2 and some from cluster 0, either directly or through an intermediate node. The composition of the graph shown in Figure 4 was done in a manual way using the information obtained from the profiling of the suspects. However, such composition could be improved through the automatic generation of the relations between suspects and the enrichment of such graph with other information obtained from `TinfoLeak` such as likes, comments, mentions,

[7] https://gephi.org

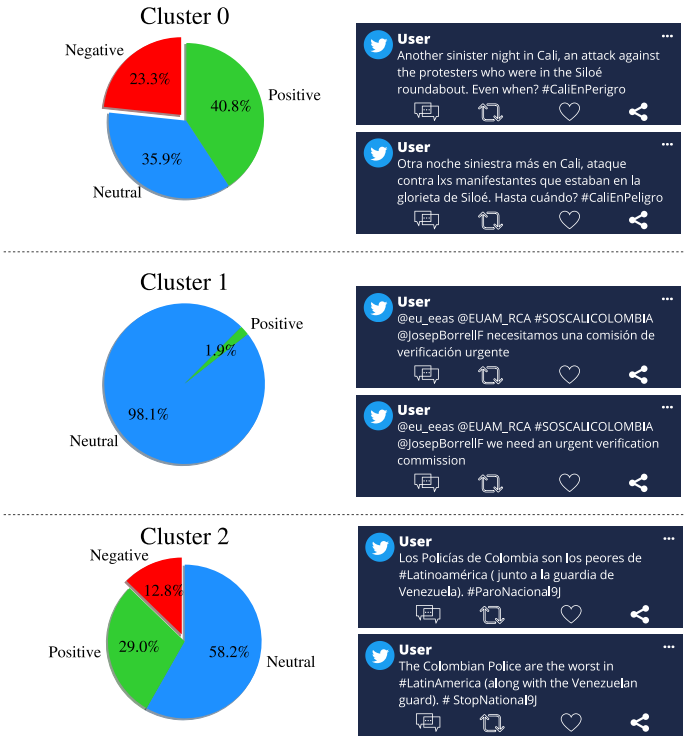

Figure 3. Clusters' sentiment polarity distribution and sample tweet for each cluster

hashtags, types of content, etc.

The methodology applied for the use case demonstrates the real power of social media in the labours of situational awareness in charge of LEAs. In particular, to spot individual actors (lonely wolfs) who generate aggressive information and have the power to influence others, and finding hostile organized communities that develop coordinated activities of manipulation and incitation.

4. A national cyberdefence strategy

Protests against governments have been witnessed in countries with robust democracy and in countries with a certain instability, regardless of the political bases exhibited by the governments. Although the protest is a legitimate population right, it may also constitute a risk for the security of the population and for the stability of the states when an agent of threat camouflages himself in those movements to generate violence and instability.

Nowadays, threat agents have explored using social networks to disseminate information and manipulate the population as a primary means [15]. This strategy is known as Hostile Social Manipulation (HSM), and it is widely used due to its impact on the population and its difficulty of being controlled by an LEA of any state [16]. On the other hand, the high amount of information disseminated during a campaign of HSM makes it tough to monitor all violent promoted actions and identify a threat agent acting behind such a campaign. Different techniques may be used as part of HSM campaigns, being disinformation operations one of the most used lately. The latter makes strong sense since false information becomes viral faster due to the sensationalism that this causes, as MIT researchers demonstrated [17].

One of the most recent cases of HSM occurred in May 2020, during the protest "*Black Lives Matter*", where the hashtag "`#DCblackout`" was used to spread false information about a generalized interruption of communications in Washington D.C. This disinformation campaign, which provoked severe incidents in Washington D.C.

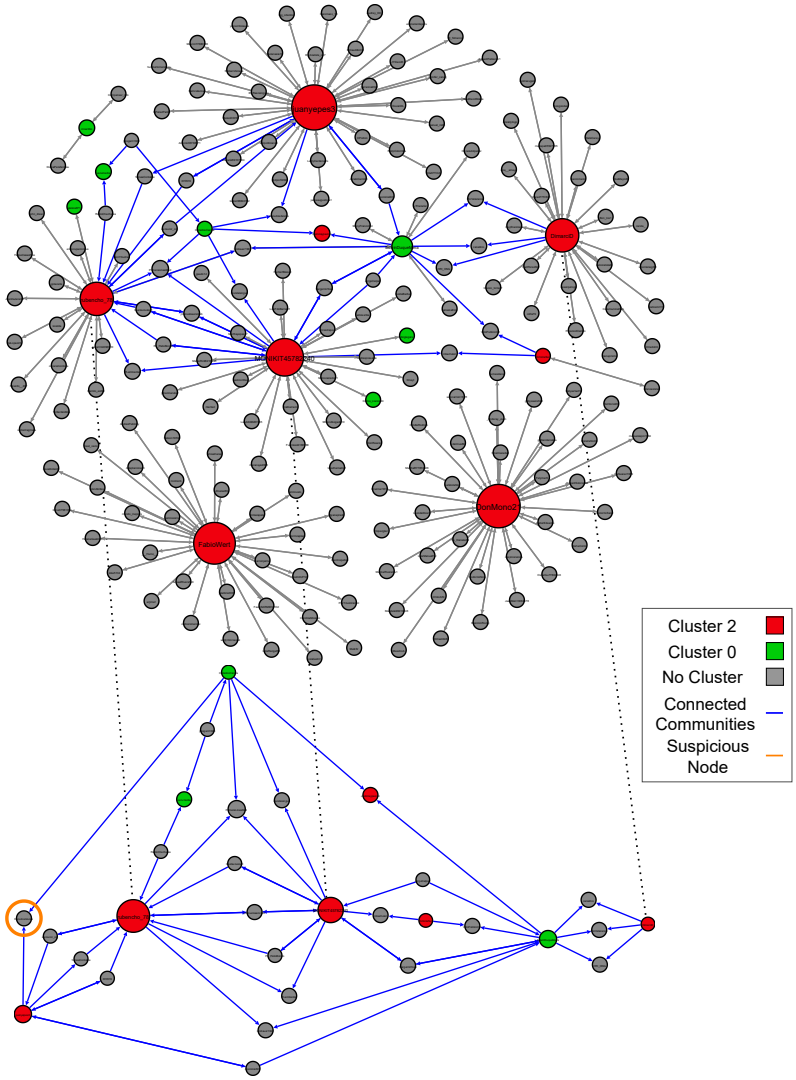

Figure 4. Graph of nodes belonging to cluster 2 showing unique and commons followers

generating multiple scenes of violence, started from an account with only three followers and became a trend in a short time.

As documented in this paper, investigating organizational and incentive tasks on social media helps to combat potential disorders. LEAs such as the Colombian Army does not see data science only as an helpful alternative but as a mandatory key piece to collect, process, and analyze information to understand threat strategies in a scalable way. Similarly, the addition of artificial intelligence (AI) techniques in an investigation led by an LEA could make the difference between success and failure. Through the adoption of solutions based on NLP, LEAs can take large amounts of messages posted in social networks, process them, compare them and identify influencing accounts during an operation of HSM.

In the paper at hand, Section 3 showed how NLP helps understand the strategy behind an HSM campaign by identifying similarities between tweets, the detection of relations between accounts, and the estimation of aggressiveness contained in the text. Other techniques such as clustering and graph representation complement in an outstanding way to extract further knowledge.

These AI-driven procedures are crucial in re-

ducing the time of analyzing social media and fueling situational awareness. Consequently, LEAs can automatically identify and monitor organizations or people in manipulation efforts for causing riots, violent movements or offensive actions on a physical level. As a result, LEAs could also deploy containment measures promptly, which may be critical in the contention of an HSM campaign.

Acknowledgments

This study was funded by the Spanish Government grants FPU18/00304 and RYC-2015-18210, co-funded by the European Social Fund. This work has also been supported by Universidad del Rosario (Colombia) through the project "IV-TFA043 - Developing Cyber Intelligence Capacities for the Prevention of Crime".

REFERENCES

1. J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, A. Vespignani, and T. Yarkoni, "Integrating explanation and prediction in computational social science," *Nature*, Jun 2021.
2. J. Pastor-Galindo, P. Nespoli, F. Gómez Mármol, and G. Martínez Pérez, "The not yet exploited goldmine of osint: Opportunities, open challenges and future trends," *IEEE Access*, vol. 8, pp. 10 282–10 304, 2020.
3. K. Müller and C. Schwarz, "Fanning the Flames of Hate: Social Media and Hate Crime," *Journal of the European Economic Association*, 2020.
4. F. Atefeh and W. Khreich, "A Survey of Techniques for Event Detection in Twitter," *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, feb 2015.
5. J. Pastor-Galindo, F. G. Mármol, and G. M. Pérez, "Nothing to hide? on the security and privacy threats beyond open data," *IEEE Internet Computing*, vol. 25, no. 4, pp. 58–66, 2021.
6. J. T. Jost, P. Barberá, R. Bonneau, M. Langer, M. Metzger, J. Nagler, J. Sterling, and J. A. Tucker, "How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks," *Political Psychology*, vol. 39, no. S1, pp. 85–118, 2018.
7. M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp, "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime," *The British Journal of Criminology*, vol. 60, no. 1, pp. 93–117, 2019.
8. D. Freelon, C. McIlwain, and M. Clark, "Quantifying the power and consequences of social media protest," *New Media & Society*, vol. 20, no. 3, pp. 990–1011, 2018.
9. J. Pastor-Galindo, M. Zago, P. Nespoli, S. L. Bernal, A. H. Celdrán, M. G. Pérez, J. A. Ruipérez-Valiente, G. M. Pérez, and F. G. Mármol, "Spotting political social bots in twitter: A use case of the 2019 spanish general election," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2156–2170, 2020.
10. L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, "Using Social Media to Predict the Future: A Systematic Literature Review," *arXiv*, no. June 2016, pp. 1–55, 2017.
11. "SoK: Hate, Harassment, and the Changing Landscape of Online Abuse." Los Alamitos, CA, USA: IEEE Computer Society, may 2021, pp. 473–493.
12. N. Alsaedi, P. Burnap, and O. Rana, "Can We Predict a Riot? Disruptive Event Detection Using Twitter," *ACM Trans. Internet Technol.*, vol. 17, no. 2, mar 2017.
13. J. Ramírez Sánchez, A. Campo-Archbold, A. Zapata Rozo, D. Díaz-López, J. Pastor-Galindo, F. Gómez Mármol, and J. Aponte Díaz, "Uncovering Cybercrimes in Social Media through Natural Language Processing," *Complexity*, vol. 2021, p. 7955637, 2021. [Online]. Available: https://doi.org/10.1155/2021/7955637
14. "Digital 2021 colombia," We Are Social Agency, Hootsuite, Tech. Rep., 2021.
15. R. Enikolopov, A. Makarin, and M. Petrova, "Social media and protest participation: Evidence from russia," *Econometrica*, vol. 88, no. 4, pp. 1479–1514, 2020.
16. M. J. Mazarr, R. M. Bauer, A. Casey, S. A. Heintz, and L. J. Matthews, "The emerging risk of virtual societal warfare: Social manipulation in a changing information environment," RAND Corporation, Tech. Rep., 2019.
17. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

Ramírez Sánchez. Julián is currently a researcher at the University of Rosario and is finishing an M.Sc. in Economics. He holds a B.Sc. in Applied Mathematics and Computer Science. His interests are related to the implementation of natural language processing algorithms. Contact him at julians.ramirez@urosario.edu.co.

Campo-Archbold. Alejandra is currently a researcher at the University of Rosario. She is pursuing a B.Sc. in Applied Mathematics and Computer Science. Alejandra is interested in the application of AI in cybersecurity. Contact her at alejan-

dra.campo@urosario.edu.co.

Zapata Rozo. Andrés is currently a researcher at the University of Rosario. He holds a B.Sc. in Applied Mathematics and Computer Science. His interests are around applying Big Data techniques in cybersecurity scenarios to help in the prevention of crimes. Contact him at andresf.zapata@urosario.edu.co.

Díaz-López. Daniel is principal professor at the School of Engineering, Science and Technology at the University of Rosario. He has a Ph.D. in Computer Science from the University of Murcia. His interests are around the use of natural language processing for the prevention of cybercrimes. Contact him at danielo.diaz@urosario.edu.co.

Pastor-Galindo. Javier is currently working toward a Ph.D. degree with the University of Murcia. His research interests focus on OSINT, security, and privacy. He received the B.Sc. and M.Sc. degrees in computer science from the University of Murcia. Contact him at javierpg@um.es.

Gómez Mármol. Félix is currently a Researcher with the Department of Information and Communications Engineering, University of Murcia. His research interests include cybersecurity, machine learning, and bioinspired algorithms. He received a Ph.D. degree in computer science from the University of Murcia. Contact him at felixgm@um.es.

Aponte Díaz. Julián is chief of the Cybernetic Development Division at the Colombian National Navy. He holds a B.Sc. in Naval Engineering and an M.Sc. in Cyberdefense. His interests are the design of secure architectures and cybersecurity policies. Contact him at julian.aponte@armada.mil.co.