Automated Statistics Extraction of Public Security Events Reported Through Microtexts on Social Networks

Flavio Ferreira da Silva*
Julio Cesar Duarte
{flavio.ferreira,duarte}@ime.eb.br
Instituto Militar de Engenharia
Rio de Janeiro, Rio de Janeiro, Brazil

Wallace Corbo Ugulino
University of Twente
Enschede, Netherlands
w.corbougulino@utwente.nl

ABSTRACT

Lately, Rio de Janeiro State has been characterized by the occurrence of successive public security events (shootings, assaults, robberies, etc.), causing great insecurity, affecting the daily lives of the population, and worrying public security agencies in the fight against crime. Although the indicators of public security events recently decreased, there is still a feeling of insecurity, while the population uses social networks to notify illegal acts that occurred in their vicinity. Although this collaboration is limited to the crimes that occurred, many published messages are difficult to interpret. Knowledge Discovery is a process of extracting data in an implicit, previously unknown, and useful way that can be applied for different purposes. In this context, Natural Language Processing is a powerful tool that allows the extraction of information from these unstructured data. This work proposes a methodology for automatic knowledge extraction, in the form of statistics related to public security events posted on social networks, particularly the ones occurred in Rio de Janeiro. The main contribution of this work is the proposal of a methodology for the construction of an Information System that allows the collection of statistics of notified public security events. In addition to this methodology, which can also be used in the construction of other Information Systems, this work contributes with a public security event recognition model that has a performance of 95%, and an available dataset that can be used to support other researches, such as: the identification of new behavior patterns, the discovery of hidden knowledge, among other fronts.

CCS CONCEPTS

• **Information systems → Web applications**; • **Computing methodologies → Information extraction**.

*Both authors contributed equally to this research.

SBSI, May 16–19, 2022, Curitiba, Brazil

ACM ISBN 978-1-4503-9698-1/22/05…$15.00
https://doi.org/10.1145/3535511.3535513

KEYWORDS

Machine Learning, Natural Language Processing, Artificial Intelligence, Public Security, Text Classification, Data Mining, Text Mining, Twitter.

ACM Reference Format:
Flavio Ferreira da Silva, Julio Cesar Duarte, and Wallace Corbo Ugulino. 2022. Automated Statistics Extraction of Public Security Events Reported Through Microtexts on Social Networks. In *XVIII Brazilian Symposium on Information Systems (SBSI), May 16–19, 2022, Curitiba, Brazil.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3535511.3535513

1 INTRODUCTION

In recent years, Rio de Janeiro State has suffered a lot with security problems, causing great insecurity, affecting the population and, at the same time, concerning public security agencies that fight against crime. Frequently, newscasts report several public security events, such as shootings, cargo theft, assaults, thefts, among others. This situation even culminated in the fact that Rio de Janeiro suffered a federal intervention in terms of security management in 2018 [4]. Although public security events indicators had a reduction in the last year [15], there is still a feeling of insecurity in the population since in mid-2020, the violence index started to grow again in Rio de Janeiro [19].

Security agencies have been committed to reducing crime in our society and social networks are a source of information that can help this process, identifying patterns of criminal behavior and hidden knowledge using data mining techniques.

Given the scenario in which the citizens live, several people make use of social networks to post about crimes that just happened, making such social media networks, especially Twitter, a collaborative tool. With that in mind, many users make use of this information to make decisions in their daily lives like, for instance, checking if there are any illicit activity happening near the transport system that takes them to work, thus being able to seek alternative routes.

Even considering this collaboration a fact regarding the crimes that occurred, many published messages are difficult to interpret. This fact is due to the way users write on social networks, not following a formalism that, according to [7], refers to the study of linguistic form, especially observing phonetics, phonology, morphology, and syntax. This way, if the posted sentence does not use simple or short words, it can lead the reader to a mistaken understanding in relation to the author's intention.

This is one of the existing problems in data mining from the internet, since, to mine texts to the point of being able to identify patterns and generate knowledge that adds value for decision making, it is necessary to use automated methods, techniques, and

tools [9]. One of these techniques is Natural Language Processing (NLP), which consists of the application of methods that enable the machine to manipulate data that systematically extract semantics from texts [16]. In Chapter VI of the Grand Research Challenges in Information Systems in Brazil 2016-2026 [3], the authors discuss how to build methodologies that can deal with citizen participation, especially electronic ones. These methodologies can be used in the implementation of systems that can help build new governmental policies and acting strategies.

Knowledge discovery consists of a non-trivial process of extracting data in an implicit, previously unknown, and potentially useful way in a database [8]. It can be applied on several fronts and for different purposes [10], however, for the purpose of this work, its use is to identify the occurrence of public security events, particularly in Rio de Janeiro State.

Regional habits end up creating terms that are informally included in the vocabulary of a language. This practice is accentuated in social networks and ends up impacting the effectiveness of NLP derived models. We can notice this mainly in public security events, where different terms like shooting, firing, and gunfire can be used interchangeably. Dasgupta et al Dasgupta et al. [6] proposed an ontology to collect and process crimes reported in US newspapers, with the aim of modeling crimes that occurred in a particular region of the country. Although our information domain has focused on Rio de Janeiro, our methodology can be applied to any region, as long as the model is trained to recognize regional terms commonly related to public security events.

The main objective of this work is to present a methodology for automatic knowledge extraction, in the form of statistics related to public security events (shootings, police operations, assaults, robberies, kidnappings, patrolling, among others), from microtexts freely published on social media networks. This is the first work, as far as we know, that explores fresh public data to generate statistics that aids public security management decision-making processes.

To summarize, the contributions of this work are three-fold. First, we propose an automated mining methodology that includes all knowledge discovery steps for microtexts processing in social networks. This methodology can be extended to other domains. In addition, we create a model that enables the recognition of public security events as named entities which can also be extended to other entities of interest. Finally, a database of collected and processed data is made available for use in future research that can help build better understanding of the behavior of public security events, techniques for discovering hidden knowledge, and event forecasting trends.

2 RELATED WORK

There are several techniques for extracting information and processing raw data from social networks, but some of them are more popular, such as NLP, Data Mining and Machine Learning (ML).

In a broader view, Dasgupta et al. [6] uses the supervised algorithm Support Vector Machines (SVM) to categorize the crimes reported in a network and extracts its entities using NLP, obtaining a performance of 94%. On the other hand, Hassanix et al. [11] provides a concise analysis of data mining applications for crimes, and, for this purpose, it analyzes more than 100 applications, making use of supervised and unsupervised algorithms to evaluate their performance.

Bendler et al. [1] proposes to analyze messages posted on Twitter related to crimes, crossing them with recorded events so that it is possible to measure the posts' effectiveness. Furthermore, it builds a model to perform predictions using the supervised SVM algorithm, achieving a performance of 71%. Likewise, Iqbal et al. [13] evaluates the prediction of two classification algorithms, Naive Bayes and Decision Trees, to observe the effectiveness of the results of both against a database of real crimes obtained from the 1990 US socioeconomic census. Thereby, it achieves a performance of 64% with Naive Bayes algorithm and 83% with Decision Tree algorithm.

While describing an architecture for information extraction systems on the web, based on NLP that is especially aimed at the exploitation of information about crimes, Pinheiro et al. [20] demonstrates the feasibility of its architecture through an implementation that provides information for a collaborative system based on the *WikiCrimes* web crime log, reporting a performance of 68%. Iriberri and Leroy [14] aimed to improve the systematic publication of crimes on the web, since, according to the authors, some existing systems do not provide support for the victims or witnesses to correctly report the crime with the relevant information to further investigation. Therefore, such work builds a system that, from some responses provided by the user, generates subsequent questions that add value to the report.

Such works have a gap which is the active and constant search for new public security events reports in online environments that are quickly updated. The present work aims to fill this gap, through the development of a methodology that allows the construction of tools capable of extracting such information in an environment as hostile as a social network that uses microtexts. To validate this methodology, a complete tool was developed that extracts and processes public security events reported about Rio de Janeiro in Twitter.

3 MINING PUBLIC SECURITY EVENTS IN SOCIAL NETWORKS

This section aims to present the operation of the generic fully automated mining methodology for collecting data from social networks that finds reports of public security events. Its data mining pipeline, as illustrated in Figure 1, is composed of several sequential actions that range from defining and parameterizing the attributes initially used, to how the processed data will be presented to the user. Flavio [5] presents more details of each of these components.

3.1 Parameterization and Initiation

The first block of our methodology is illustrated in Figure 1, as the "Parameterization and Initiation" Lane. Its first task is "Parameterize Search Data" whose purpose is to instantiate the objects necessary for the automated mining methodology to work correctly inside the knowledge discovery process. The activities of this task carry out all the initial configurations and setups so that the automated mining methodology works.

Following the step by step in Figure 1, the next task to be performed is "Submit Search Engine" whose function is to collect data

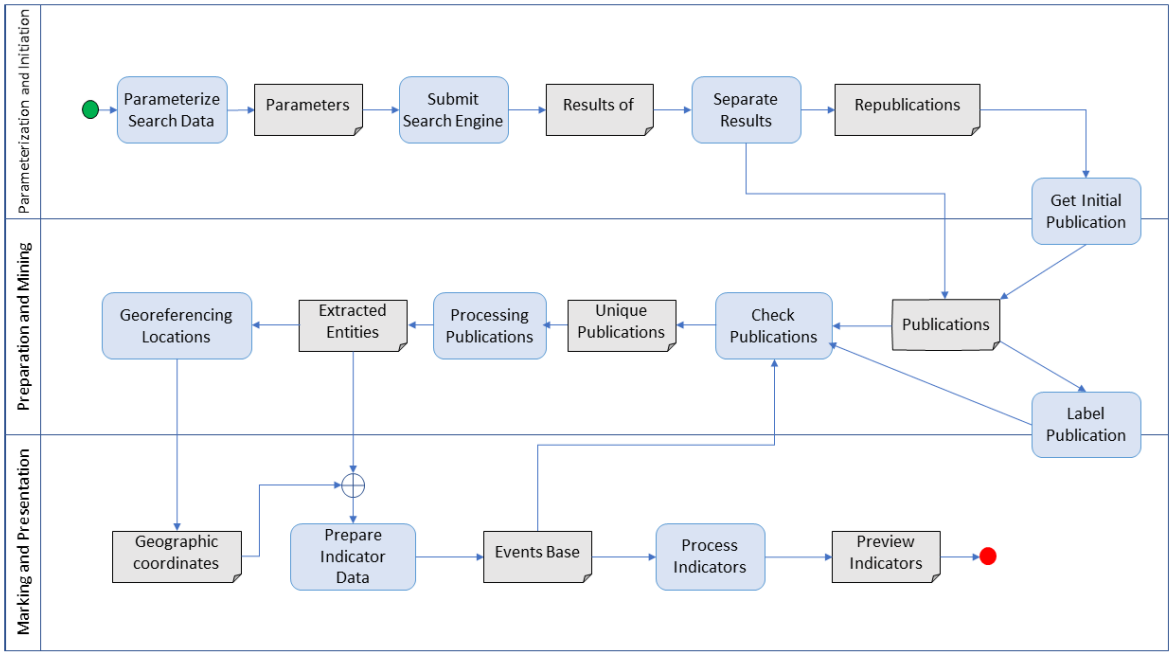

Figure 1: Automated Mining Methodology Blocks and Tasks. Adapted from Flavio [5].

from social networks that are inputs for the entire process of identifying patterns and discovering knowledge.

In order to improve the posts' collection performance, we store data in two different repositories, so that the whole collection process is more robust, thus, ensuring its availability, integrity, reliability and authenticity of the data. The "Separate Results" task has activities that split the collected data into two named groups: original messages and re-posted messages. This action was taken so that the automated mining methodology primarily processes only the original messages.

Other tasks are performed in this methodology but are not part of any block. They are represented in Figure 1 by the "Get Initial Publication" and "Label Publications" tasks. The first task is intended to ensure the integrity of the database, preventing a reposted message from not being associated with an original message. This ensures a hierarchical relationship between messages so that the automated mining methodology checks messages that meet this criterion and triggers the process to obtain the original one.

The "Label Publications" task is also not part of any block and does not need to be performed every time. It may be triggered only, depending on performance of the entities recognition systems. The activities of this task are based on learning, training, and evaluating the algorithms, to make the entity recognition process more efficient.

3.2 Preparation and Mining

All data mining tasks are important, since the results of one aid others, so the "Preparation and Mining" block is intended to treat and process data coming from the "Parameterization and Initiation" block to identify hidden patterns and knowledge discover. It is patent that a lot of data retrieved from social networks comes with noise issues, duplicates and even inconsistencies. Therefore, in addition to prepare data for the application of mining algorithms, this block is also responsible for the cleaning, consolidation, integration, and separation of data which is performed by the "Check Publications" task. It was introduced into the mining process because one of the characteristics of data collected from the Internet is the unstructured way in which information is provided. Thereby, the mining process becomes more complex since there is no formal writing pattern in social networks.

The next task performed in this block is the core of every automated mining methodology and is represented in Figure 1 by the "Processing Publications", which is responsible for the entities extraction and site logging. With the speed with which information is generated on social networks, it is increasingly difficult to achieve acceptable efficiency in data mining, given its complexity. Another important point in this same context is the way in which messages are elaborated, with no formalism, neither in spelling nor in semantics. For example, a user can make a publication combining texts, emojis and symbols, assigning the message a semantic value that a simple system would not understand, thereby reducing the effectiveness of the entity extraction process.

In this work, entity recognition is conducted through two models that use ML algorithms. One of them is native to the interface used and the other was created specially to meet the specifics of this work. This new model was, then, coupled with the automated mining methodology that makes use of this expedient to extract a new public security event entity. The creation of this model was necessary because this new entity is not automatically recognized by the native model.

The ML algorithms used to build the new entity recognizer use a sophisticated word embedding strategy with subwords and Bloom embeddings' capabilities fed into a deep neural network with residual connections [18].

This new entity was labeled as Public Security Event and the terms that represent them are assault, stray bullet, police operation, robbery, cargo theft, shooting, among others. The posted messages basically have four entities that need identification, they are: the public security event itself, its location, date, and time. Other entities are also identified in this task, but they are only used in the "Marking and Presentation" block.

When the entity extraction sub-task is triggered, it processes all original messages that were stored in the database. For each message stored, the automated mining methodology applies the native model for recognizing the location, date, and time entities, and, for the public security event entity, it applies the new generated model that also recognizes secondary entities (author, followers, shares, likes, and others) which can be used in the "Marking and Presentation" block.

The next task "Georeferencing Locations" fetches the geographic coordinates of each location identified in the "Processing Publications" task and stores them in the database. It also separates locations where latitudes and longitudes were retrieved from those that could not be found. First, this task accesses the database with the locations that were successfully identified and triggers the model to connect with the geolocation interface to start identifying their coordinates. Once these activities are completed, the "Georeferencing Locations" task accesses a geolocation service and submits the locations to identify their latitudes and longitudes.

3.3 Marking and Presentation

Finally, the last block of the automated mining methodology is the "Marking and Presentation" block which is represented in Figure 1 by the "Prepare Indicator Data" and "Process Indicators" tasks. Its purpose is to allow users to visualize the results in the form of indicators to support a possible decision-making process. According to [2], data only adds value to the organization if there is a possibility that this data undergoes transformation. The first task, "Prepare Indicator Data", has the purpose of preparing and filtering the processed data, so that the "Process Indicators" task assembles metrics and statistics, making them available for the analysis and the decision-making processing.

4 IMPLEMENTATION AND RESULTS

This section aims to present the implemented prototype of the automated mining methodology in the context of the public security events processing. As an use case, we applied the methodology to events that took place in Rio de Janeiro and were reported through the Twitter social network. The implementation process divides the methodology into four layers, which are then subdivided into tasks that provide greater clarity about the activities that are performed within each of the layers.

These layers represent the implementation of the methodology proposed in this work and comprise the specifics and limitations related to the use case being implemented here. Figure 2 illustrates how the layers of the automated mining methodology are implemented. In addition to presenting the technical requirements for the implementation, this chapter presents a quantitative assessment of the results of the processing of each of these layers.

The methodology for automated mining was implemented using a bot developed in the Python programming language that persists the information obtained in a PostgreSQL database. The bot is built, configured, parameterized, and loaded inside the Recovery layer.

4.1 Recovery Layer

The Recovery layer is implemented comprising two tasks, "Data Retrieval" and "Data Storage", which characterize the entire beginning of the process of the automated mining methodology. When the process is started by the "Data Retrieval" task, the bot accesses the Twitter API and starts retrieving the tweets.

The bot retrieves the tweets in its two available formats, .CSV and .JSON. This action is taken due to a concern of processing time for storage and because of the greater variety of information available in former format. After completing the retrieval process, in a second moment (offline), data is stored in the database.

4.2 Transformation Layer

The "Data Cleaning" and "Data Separation" tasks were implemented inside the Transformation layer and their purpose is to prepare data for the application of the NLP techniques. The main idea here is to process only a part of the information from the posted tweets, storing the remaining data in a repository for possible future use.

The "Data Cleaning" task removes all information that is not used in the processing layer, since when messages are collected on Twitter, they come with a lot of "unusable" data that, in many cases, are included by their authors. For example: emojis, photos, videos, mentions to other accounts, ellipses, animated GIFs, among others. Such elements are removed to improve the effectiveness of the knowledge discovery process.

After removing these elements, the bot stores the tweets in a new database table, containing the tweet text, date, time, tweet ID, retweet ID, author, author photo, number of followers, number of retweets and geolocation. Table 1 shows an example of a tweet before and after the "Data Cleaning" task. In the example of Table 1, the tweet, which is written in Portuguese, means "There is a police operation at São José Park in Belford Roxo, Baixada Fluminense (7:28)".

Table 1: Tweet before and after cleaning.

Original Tweet	Processed Tweet
Há operação policial no Parque São José em Belford Roxo, na Baixada Fluminense, às 07:28 #RiscodeTirosRJ #FogoCruzadoRJ https://t.co/6GmwWiSY9X	Há operação policial no Parque São José em Belford Roxo Baixada Fluminense às 07:28

4.3 Mining Layer

The tasks where the application and execution of NLP techniques occur are implemented by the Mining layer. It consists of four tasks

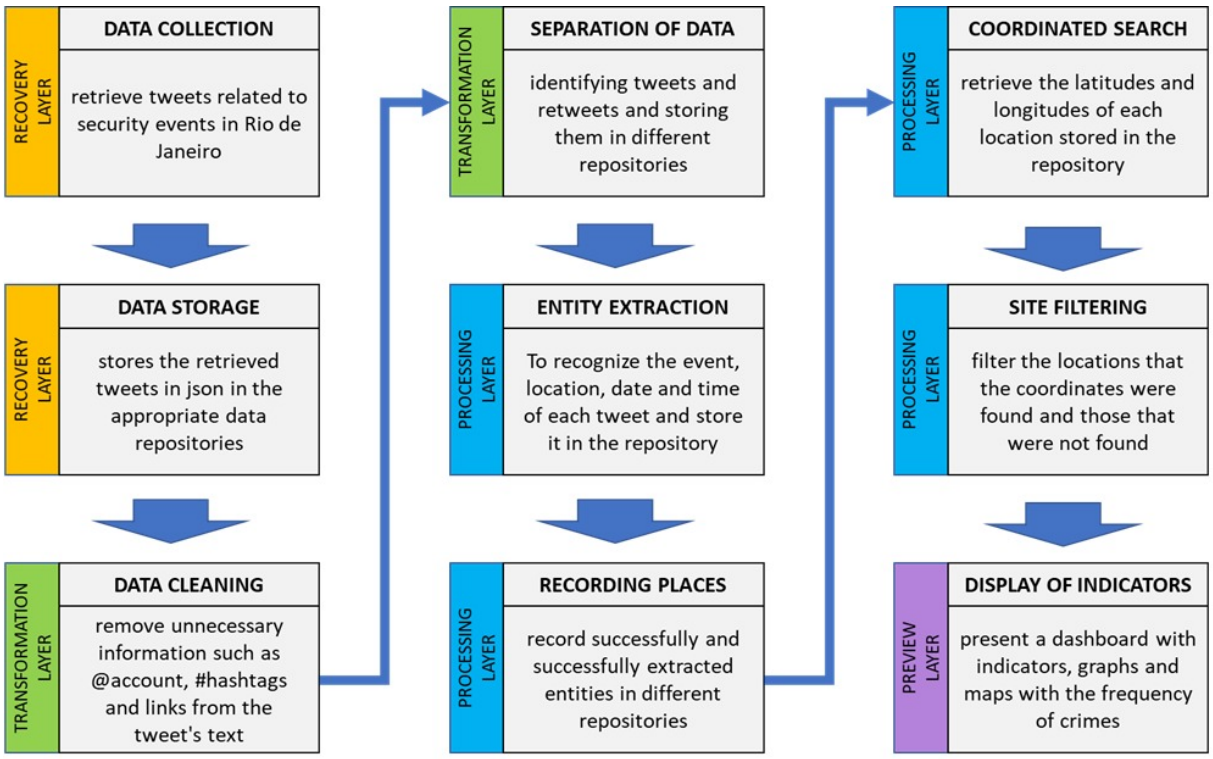

Figure 2: Automated Mining Methodology Flow. Adapted from Flavio [5].

called "Entities Extraction", "Site Recording", "Coordinates Search" and "Sites Filtering". Each of these tasks plays a significant role in data mining and hence knowledge discovery. A key point in this work is the use of the Spacy library [18] for the dynamics of extracting entities. Spacy allows the creation of statistical models for the training and recognition of entities.

Tables 2 and 3 illustrate the training and extraction, respectively, of these entities, where the numerical values indicate the position of the starting and ending characters of the entity being recognized.

Table 2: Extract from the examples of the training of the public security event entity.

Training of the public security event
('Tiros atinge policial dentro da UPP','ent.':[(0, 5, 'EVENT')])
('Suspeito de assalto é morto na rua','ent.':[(11,19,'EVENT')])
('Mulher é atingida por bala perdida','ent.':[(22,33,'EVENT')])

In Table 2, the examples are, in English, "Shots hit police inside the UPP", "Suspect of robbery is killed in the street", and "Woman is hit by stray bullet". In Table 3, the example means "Shots in the Himalaia Hill, in Chapadão Complex, Anchieta".

We collected more than 12,000 tweets during the year of 2020 and the dataset is made available[1] for download in a CSV format.

[1]https://colossos29.com.br/download/

Table 3: Extracting entities from a tweet.

Entity	Instance
Tweet	22/04/2020, 12:29:00, Tiros no morro do Himalaia, no complexo do Chapadão em Anchieta.
Event	Tiros
Site	[0] Himalaia; [1] Chapadão; [2] Anchieta;
Date	Apr, 22nd 2020
Time	12:29:00 PM

Tables 4 and 5 present a sample of the last seven weeks of data collection with their respective processing results in terms of model performance. Table 4 presents the hit rates and processing times in terms of recognizing the Public Security Event and Location entities. Table 5, on the other hand, presents the hit rates in retrieving their coordinates - latitudes and longitudes - of each Site entity. The entire evaluation was performed on a dataset not used in the entity recognizer training.

Data listed in Table 4 show the performance, both in terms of time and accuracy in the processing of the entity extraction step. The DMD column, Data Mining Duration, presents the time taken for the bot to identify the Public Security Event and Location entities in all tweets. The column CERR, Correct Entity Recognition Rate, shows the accuracy in recognizing public security event entities, as well as CSRR, Correct Site Recognition Rate, that shows its accuracy for the sites. The items listed in Table 5, on the other

Table 4: Data mining results.

Week	# Tweets	DMD	CERR	CSRR
1	423	4.38 seg	99.30%	94.37%
2	478	4.43 seg	97.16%	97.87%
3	412	8.46 seg	89.64%	97.93%
4	128	2.73 seg	97.96%	95.91%
5	130	2.80 seg	97.50%	93.75%
6	131	2.72 seg	96.72%	93.44%
7	226	2.75 seg	86.25%	91.25%

Table 5: Results for Site Coordinates Identification.

Week	# Tweets	CSD	CCRR
1	423	28.92 seg	82.11%
2	478	34.86 seg	78.15%
3	412	32.95 seg	61.60%
4	128	50.28 seg	88.89%
5	130	50.70 seg	100.00%
6	131	50.86 seg	92.59%
7	226	50.74 seg	85.71%

hand, present data obtained in the step that recovers the geographic coordinates of each identified location. The columns CSD, Coordinate Search Duration and CCRR - Correct Coordinate Recognition Rate – represent, respectively, the processing time and accuracy of the bot in the execution of the process of searching the precise coordinates for the found locations.

4.4 Preview Layer

The Preview layer is the phase in which data processed in the previous layers is consolidated in the form of measures and metrics, so that the end user can consume this information to support their decision making. It can also support the construction process of strategic guidelines for security agencies. The "Indicators Display" task is designed to present data to the user in the form of online indicators (measures and metrics), and it was implemented using two of the most used frameworks in Python.

In general, this layer uses tables and maps that show event frequency distributions to present public security events and their highest incidence regions. This aids in behavioral analysis of events and planning for operations that can lead to a reduction in event growth.

Django was used for the web application development, and Folium is the library used to dynamically build charts and maps. Such libraries have different purposes, but they complement each other in the context of this methodology. According to Python-Visualization [12], Django is a framework for developing web applications that was born in 2003 to meet a journalistic demand. It allows building a dynamic website quickly, enabling the separation of the implementation of the algorithms from the application design itself.

The Folium library [17] is used to create maps with the identified public security events, which are the most important step in the "Indicators Display" stage, as it allows the user to have a macro view of how public security events are distributed geographically. In this way, different types of maps can be generated, but two of them are more prominent.

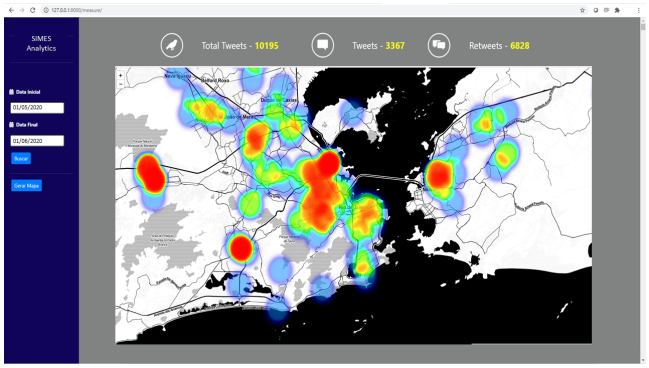

Figure 3: Heat map of some public security events reported in Rio de Janeiro.

The "Heat Map" allows the presentation of areas with the highest frequency of public security events, as illustrated in 3. The "Points Map", on the other hand, allows the visualization of the precise locations where the public security events occurred, or at least, where they were reported, as well, as detailed information of the reported event, including the tweet itself, for its validation, as illustrated in 4.

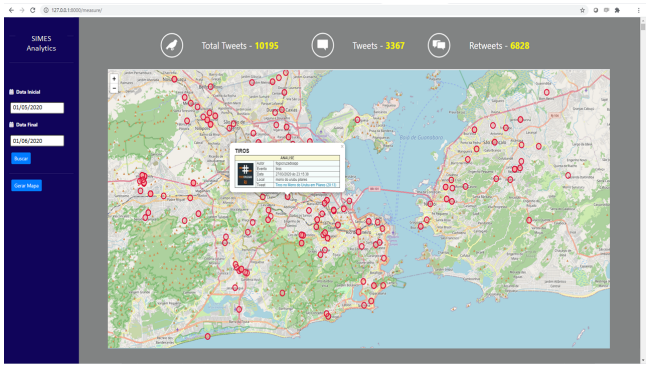

Figure 4: Points map of some public security events reported in Rio de Janeiro.

5 CONCLUSIONS

The advent of social networks is a fact, and their use tends to increase every day, generating more information that needs to be processed. Hence the need for automated methodologies that can process and filter data to present the user with only the essential information needed for their decision-making process. Another highlight that needs to be considered, as a relevant piece in this context, is the participation of the population in social networks, contributing to society in reducing crime by reporting public security events on their social networks. It is worth remembering that

these data supported the entire formulation of this work. We can see this act by the population as a way to report a public security event in a collaborative way.

As direct contributions of this work, we can mention three major aspects. The first is the proposal of a new methodology that enables the execution of all phases for knowledge discovery, as well as the implementation of extractors with specific natural language processing techniques that recognize public security events. The second contribution is the creation of a model for the recognition of named public security event entities, which can also be used in areas other than Rio de Janeiro. Finally, the last contribution is the construction of a dataset with collected and processed data to be used in future research.

The system presented and developed here can be used in conjunction with other public security event reporting tools already being used by Public Agencies, complementing the reported events, by other means, or aggregating in order to provide more accurate statistics in relation to what happens daily in a particular place.

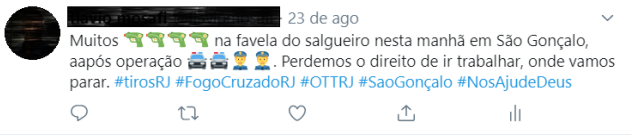

Figure 5: Example of published tweet combining text and emoji.

In addition to the contributions cited here, future work may focus on improving the entity extractor to use other information such as emojis, as shown in Figure 5, where the text reads, in English, "Many shots in Salgueiro community this morning in São Gonçalo, after a police operation. We lost the right to work, where we are going to stop.". Such emojis can be used to express an emotional condition or part of the tweet context.

It is also expected that in the future this tool may be used in conjunction with the application of prediction techniques of machine learning algorithms to identify behavioral patterns related to these public security events.

Finally, this work showed several possibilities where the methodology can provide useful information to security agencies, being one additional component to fight crime with intelligence. It is clear from the results presented that there are places, behaviors, and specific periods where public security events occur more frequently, evidencing a systematic of these acts. Therefore, security agencies can make use of this expedient to better understand criminal acts, in order to neutralize them, providing an increasingly safe environment for society.

ACKNOWLEDGMENTS

This work was partially supported by national funds through FINEP, Financiadora de Estudos e Projetos and FAPEB, Fundação de Apoio à Pesquisa, Desenvolvimento e Inovação do Exército Brasileiro, under project "Sistema de Sistemas de Comando e Controle" with reference nº 2904/20 under contract nº 01.20.0272.00.

REFERENCES

[1] Johannes Bendler, Tobias Brandt, Sebastian Wagner, and Dirk Neumann. 2014. Investigating crime-to-twitter relationships in urban environments - facilitating a virtual neighborhood watch. In *Proceedings of the European Conference on Information Systems* (22 ed.). ECIS, Israel, 1–15.

[2] Eduardo Bezerra, Emanuel Passos, and Ronaldo Goldschmidt. 2015. *Data Mining: Conceitos, Técnicas, Algoritmos, Orientações E Aplicações.* Campus, Rio de Janeiro, Brazil. 296 pages.

[3] C. Boscarioli, R. M. Araújo, and R. S. P. Maciel. 2017. *I GranDSI-BR – Grand Research Challenges in Information Systems in Brazil 2016-2026.* Special Committee on Information Systems (CE-SI)., Brazilian Computer Society (SBC).

[4] Presidência da República. 2020. DECRETO Nº 9.288, DE 16 DE FEVEREIRO DE 2018. http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Decreto/D9288.htm#textoimpressao 16 nov. de 2020.

[5] Flavio Ferreira da Silva. 2020. *Metodologia para a Extração Automatizada de Estatísticas Relacionadas a Eventos de Segurança de Microtextos das Redes Sociais.* Master's thesis. Instituto Militar de Engenharia.

[6] Tirthankar Dasgupta, Abrir Naskar, Rupsa Saha, and Lipika Dey. 2017. Crime-Profiler: Crime Information Extraction and Visualization from News Media. In *Proceedings of the International Conference on Web Intelligence* (1 ed.) *(WI-17, 9).* ACM, New York, 549–549.

[7] Mike Dillinger. 1991. Modeling message diffusion in epidemical DTN. *Ad Hoc Networks* 16, 2 (1991), 197–209.

[8] Usama Fayyad, Gregory Piatetsky Shapiro, and Padhraic Smyth. 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine* 17, 3 (1996), 37–54.

[9] Ronen Feldman and James Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data* (1 ed.). Cambridge University Press, New York. 476 pages.

[10] Jiawei Han, Micheline Kamber, and Jian Pie. 2012. *Data Mining: concepts and techiques* (3 ed.). Elsevier, United States of America. 476 pages.

[11] Hossein Hassanix, Xu Huang, Emmanuel S Silva, and Mansi Ghodsi. 2016. A Review of Data Mining Applications in Crime. *Stat. Anal. Data Min.* 9 (2016), 139–154. https://doi.org/10.1002/sam.11312 12 nov de 2018.

[12] Adrian Holovaty and Jacob Kaplan-Moss. 2008. *The Definitive Guide to Django: Web Development Done Right* (2 ed.). Apress, New York. 433 pages.

[13] Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Payam Hassany Shariat Panahy, and Nasim Khanahmadliravi. 2013. An Experimental Study of Classification Algorithms for Crime Prediction. *Indian Journal of Science and Technology* 6 (2013), 4219–4225. 3 mar de 2013.

[14] Alicia Iriberri and Gondy Leroy. 2007. Natural Language Processing and e-Government: Extracting Reusable Crime Report Information. In *2007 IEEE International Conference on Information Reuse and Integration.* IEEE, Las Vegas, Nevada, USA, 221–226. https://doi.org/10.1109/IRI.2007.4296624

[15] Instituto Segurança Publica ISP. 2020. Instituto de Segurança Publica divulga dados do primeiro semestre. https://www.isp.rj.gov.br:4431/Noticias.asp?ident=441 16 jul. de 2020.

[16] Aydano Machado. 2010. Mineração de Texto em Redes Sociais Aplicada à Educação a Distância. *Mineração de dados* 530 (2010). http://pead.ucpel.tche.br/revistas/index.php/colabora/article/view/132 07 out. de 2018.

[17] Python Data Leaflet Maps. 2020. folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the Leaflet.js library. Manipulate your data in Python, then visualize it in a Leaflet map via folium. https://pypi.org/project/folium/ 13 jun. de 2020.

[18] Honnibal Matthew and Montani Ines. 2020. spaCy 2 Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/. [Online; accessed 25-Abril-2020].

[19] O Globo. 2020. Mesmo em meio à pandemia, mortes violentas crescem 7% no primeiro semestre. Levantamento do Fórum Brasileiro de Segurança Pública aponta que redução de circulação de pessoas nas ruas não impediu aumento nos números. https://oglobo.globo.com/brasil/mesmo-com-pandemia-mortes-violentas-crescem-7-no-primeiro-semestre-1-24699633 19 out. de 2020.

[20] Vládia Pinheiro, Vasco Furtado, Tarcisio Pequeno, and Douglas Nogueira. 2010. Natural Language Processing based on Semantic inferentialism for extracting crime information from text. In *International Conference on Intelligence and Security Informatics* (1 ed.). IEEE, Canada, 19–24.