Journal of Information Systems Engineering and Management
2020, 5(4), em0126
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

MODESTUM
OPEN ACCESS

An Empirical Study of the Perception of Criminality through Analysis of Newspapers Online

Manuel Saldaña [1,2]*

[1] Faculty of Engineering and Architecture, Universidad Arturo Prat, Almirante Juan José Latorre 2901, Antofagasta 1244260, CHILE
[2] Department of Computing and Systems Engineering, Universidad Católica de Norte, Angamos 0610, Antofagasta 1270709, CHILE
***Corresponding Author:** masaldana@unap.cl

Citation: Saldaña, M. (2020). An Empirical Study of the Perception of Criminality through Analysis of Newspapers Online. *Journal of Information Systems Engineering and Management, 5*(4), em0126. https://doi.org/10.29333/jisem/8492

ARTICLE INFO

Published: 30 Aug. 2020

ABSTRACT

Crime analysis represents a great challenge to law enforcement considering that the sources to use for generating intelligence are diverse in content and/or structure. However, in recent years, techniques such as natural language processing, a field of computing, artificial intelligence and linguistics have been developed that allow to study the interactions between computers and human language, and that in turn can be used effectively in the analysis of large amounts of texts and in the subsequent derivation of interesting analytical results. This paper presents a model for analysis of criminal events from online newspapers, identifying the areas with the highest crime rates through the detection and geographical mapping of critical points and the analysis of the nature of the criminal event. The evaluation of the proposed model to estimate the perception of crime in the domain of the proposed communes indicates that it is efficient in categorizing the news and the nature of these (validated by the performance indicators).

Keywords: criminal analysis, geographical mapping, perception of criminality, web mining

INTRODUCTION

Crime analysis has existed since before the time of Sherlock Holmes, when the public was introduced to the analysis of facts and information to solve the crimes of the time, through the use of the scientific method, logic, and the powers of observation and deduction (Petherick, 2015). Because crime analysis is a very broad term, it encompasses a variety of different practices with different focuses and outcomes. This is discussed by Boba (2016), where it indicates that the central focus of crime analysis is the study of crime and disorder, problems and information related to the nature of incidents, offenders, and victims or targets of these problems. Crime analysts also study other police-related operational issues, such as staffing needs and areas of police service. Thus, even though this discipline is crime analysis, in practice it includes much more than just the examination of crime. Then, the criminal analysis is understood as the study of all elements involved in a crime. Searching, analyzing and linking criminal data it is possible to find meaningful information that allows order and security forces to clarify crimes, arrest criminals and prevent the occurrence of future criminal events (Nokhbeh Zaeem et al., 2017; Po and and, 2018).

The methodologies of criminal analysis are currently a priority area of work in the prevention of crime worldwide (Boba, 2001; European Commission, 2011; O'Shea and Nicholls, 2002; Stenton, 2006; Vellani, 2010), developing investigations that help to identify the skills or qualities that efficient crime analysts must have (Evans and Kebbell, 2012). Generation intelligence between data and data sets, for example, related to incidents and criminal acts and the patterns of occurrence in a certain place, may contribute to improve the response to demands or requirements of security or justice (Marinescu and Balica, 2018), whose benefits and/or disadvantages can be found in the literature.

The state of the art indicates the adoption of technologies, the exploration of information, and the impact of these on police practices on the ground, through the organizational and cultural integration of crime analysis, the technological support of analytical practices and the incorporation crime analysis for police practices (Sanders and Condon, 2017), inclusion of geographic features in machine learning algorithms for the prediction of crime based on the network by incorporating a criminal rack (Lin et al., 2018), development of algorithms to find patterns of money laundering criminals (Badal-Valero et al., 2018), development of fuzzy clustering algorithms K-means to obtain criminal critical points, indicating locations with high crime incidence, together with formal concept analysis used to extract visual models that describe patterns that characterize criminal activities (De Farias et al., 2018), analysis of criminal network activities through the application of deep reinforcement learning, applied to the development of a prediction model of hidden links of criminal networks (Lim et al., 2019a, 2019b) and link prediction model that incorporates a merger of metadata with a criminal data set that evolves over time (Lim et al., 2020), fitting statistical models and

machine learning for predicting recidivism (Dressel and Farid, 2018; Tollenaar and Van Der Heijden, 2009, 2019), the identification of criminal organizations from social network structures, through the evaluation of common metrics for social network analysis, modeling with decision trees and frequency analysis of network motives (Cesur et al. 2017; Çinar et al., 2019), detection and prevention of fraudulent activities related to financial institutions (Makki et al., 2019), visual content generation and natural language processing as a method of teaching skills in police academies (De Sousa Netto et al., 2019), serial crime detection by linking the "modus operandi" (M.O.) and the information of the criminal process, using a natural language processing method to extract the characteristics of the action and object of the criminal process, in addition to an information entropy method to weigh the similarity of the action and the characteristics of the object to obtain the comprehensive similarity of the penal process of criminals (Li and Qi, 2019) and detection of the intention of potential criminal acts through social networks through the generation of ontologies (Saldaña et al., 2019; Yang et al., 2008) (mainly of a specific slang) and machine learning techniques (de Mendonça et al., 2020), among others.

On the other hand, Meijer and Wessels (2019) conclude that the current thrust of predictive policing initiatives is based on convincing arguments and anecdotal evidence rather than on systematic empirical research, and urge the research community to do independent tests of both positive and negative expectations to generate an evidence base for predictive policing. While that Belur and Johnson (2018) suggest that while crime analysis is acknowledged as being central to the business of everyday policing, the capabilities are being underutilization.

Considering the above, there is a study that not only considers crime based on the geographical area, but also considers the population's perception of crime according to zones within a city, is not exhaustively developed in the literature. However, the work carried out by Saldaña et al. (2020) develop a mapping of the perception of robbery crimes analyzing a mass media, such as articles of online newspapers. Then, to study the perception of the community with respect to criminal events, an analysis of news extracted from a variety of online newspapers is developed, and then, news are processed with a text analysis tool, it proceeding to the recognition of entities of location of the commune where the robbery event took place and the nature of the event.

The development of this research includes an introduction to web mining, works related to the extraction of information from online documents, the methodology and implementation proposed, results, and finally conclusions and future works. The practical applications is due to the importance of knowledge of variables such as the amount of crimes and the nature of them when making decisions about where to move house or which places to visit when traveling, and at what times, or what places to avoid, etc. For example, someone could be interested to compare different cities or zones inside a city according to criminality or compare different neighborhoods in a city to choose a safer one, or a traveler could be interested in know what parts of a city he should avoid. Currently, this information is not available for everyone, but it can be generated from the analysis of newspaper articles online of a region. Another potential interested could be the town halls and order forces, which can use this tool to identify critical crime points and then use the appropriate control mechanisms.

MATERIALS AND METHODS

Theoretical Framework

Web Mining is the process of data mining techniques to automatically discover and extract information and useful knowledge from web. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns, for example, the analysis of traffic, the most popular contents, or demographic data of the visits. While web data mining is referred to the generation of intelligence from data retrieved from the web, through of the application of data mining techniques (Kumar, 2015; Sidana and Aggarwal, 2017), also can be defined as the processes responsible of the discovery of information that does not exist explicitly (Liu, 2011).

Text mining, on the other hand, refers to the process of discovering interesting knowledge from text documents (Talib et al., 2016), through text processing to form new facts and hypotheses, which can be further explored with other data mining algorithms (Vishal Gupta, 2009). The selection of features aims focusing only on relevant and informative data for use in text mining, and some of the topics of text mining include feature extraction, text categorization, clustering, trend analysis, association mining, and visualization (Hotho et al., 2005; Kaushik and Naithani, 2016). Text Mining process unstructured information extract significant numerical indexes from the text and, in this way, make the information contained in the text accessible to the various data mining algorithms. With text mining is possible analyze words, groups of words used in documents, etc., or analyze documents and determine similarities between them or how they relate to other variables of interest, in other words, text mining will "convert the text into numbers" (meaningful indexes) (Justicia de la Torre et al., 2018).

Finally, web mining has to be used to automatically discover and extract information from web-related data sources, such as documents (Hammouda and Kamel, 2004), records, services and user profiles, and although standard data mining methods can be applied for Mining in the Web, some algorithms need to be developed and applied for the processing of information based on the Web, such as the development of ontologies as a method of representing the recovered information (Li and Zhong, 2004).

Related Works

Between 2002 and 2005, product of the Cornwall Crime Surveys (CCS) in a rural county of England, Mawby (2004) development an exploratory model of the occurrence of crimes and places where they were carried out. Abdul Jalil, Mohd and Mohamad Noor (2017), on the other hand, present a comparative study on correlation and information gain algorithms to evaluate and produce subsets of criminal characteristics, identifying a subset of attributes and classify the crimes into different categories, predicting the category of crime and directly support decision-making in crime prevention systems. A relevant point to consider is the

inclusion of the information obtained from the mass media in the network, such as social networks, online newspapers or mass-use websites as YouTube (Adnan et al., 2011; Kahya-Özyirmidokuz, 2016; Pinto et al., 2017; Song et al., 2016), platforms that is becoming the largest source of public access data in the world and that make extracting useful information and knowledge a fascinating and challenging task (Liu, 2011).

The exponential increase in online social media allows users around the world to share and communicate information and ideas freely through the internet (Dowerah Baruah, 2012), becoming in a dominant communication tool and that has been used as a communication channel in several events, how for example, “The Arab Spring” and BOSTON’S attack, etc. (Alami and Elbeqqali, 2015). To develop useful profiles of different cybercriminals, text mining techniques are an effective way to detect and predict criminal activities in microblog publications that consider the problems of data scarcity and semantic gap (Gerber, 2014), developing methodologies to apply complex networks in the analysis of criminality disseminated within criminal geographic areas within a city (Spadon et al., 2017), using tools such as ontologies for analyze inter-gang relationships, linking criminals to certain criminal organizations and relationships between them (Vishwakarma and Shankar, 2014).

The geographic mapping of crimes can be carried out by extracting relevant information from unstructured data from online newspaper articles (Saldaña et al., 2020). The automatic extraction of still hidden public information available in newspaper articles can indicate the frequencies of crimes in certain sectors of a city by identifying locations using recognition algorithms of named entities (Arulanandam et al., 2014). Newspapers are a source of authentic and timely information (mostly), contain information about crimes, accidents, cultural and sports events, among others, and despite having this valuable information available, the use of it for the generation of intelligence has not yet been generalized.

Data mining is a powerful tool that can be used effectively to analyze and derive important analytical results (Sathyadevan et al., 2014). Jayaweera et al. (Jayaweera et al., 2015) proposed a web-based system that includes techniques for analyzing crimes such as the detection of critical points, the comparison of crimes and visualization of patterns of delinquency. While that Chen et al. (Chen et al., 2004), several data mining techniques are exposed, both for local security applications and for national security applications (framed in the post attack context September 11). Lim et al. (Lim et al., 2018) developed an analysis of several cases of big data use in cities around the world, cities that try to become smart cities and where the use of urban big data plays a preponderant role, identifying the challenges of this transformation and the directions for its implementation (Pan et al., 2016), cities that are being flooded with data (Andrienko et al., 2016). From the perspective of data science, the data that emerge from smart cities give rise to many challenges that constitute a new interdisciplinary field of research and that have the potential to be a useful tool in the study and generation of patterns criminals and crime prevention plans (Lim et al., 2018), however, in a hyper-connected society, the concept of privacy would become a paradox, since fundamental human rights must be protected and the collection and analysis of data should provide a conscious approach to privacy (Pan et al., 2016; Rouvroy, 2016).

ARCHITECTURE OF THE PROPOSED MODEL

The objective of this research work is to identify theft locations within the body of the news and the nature of criminal acts, to create intelligence from public information. This section shows a high-level diagram (see **Figure 1**) for the information extraction

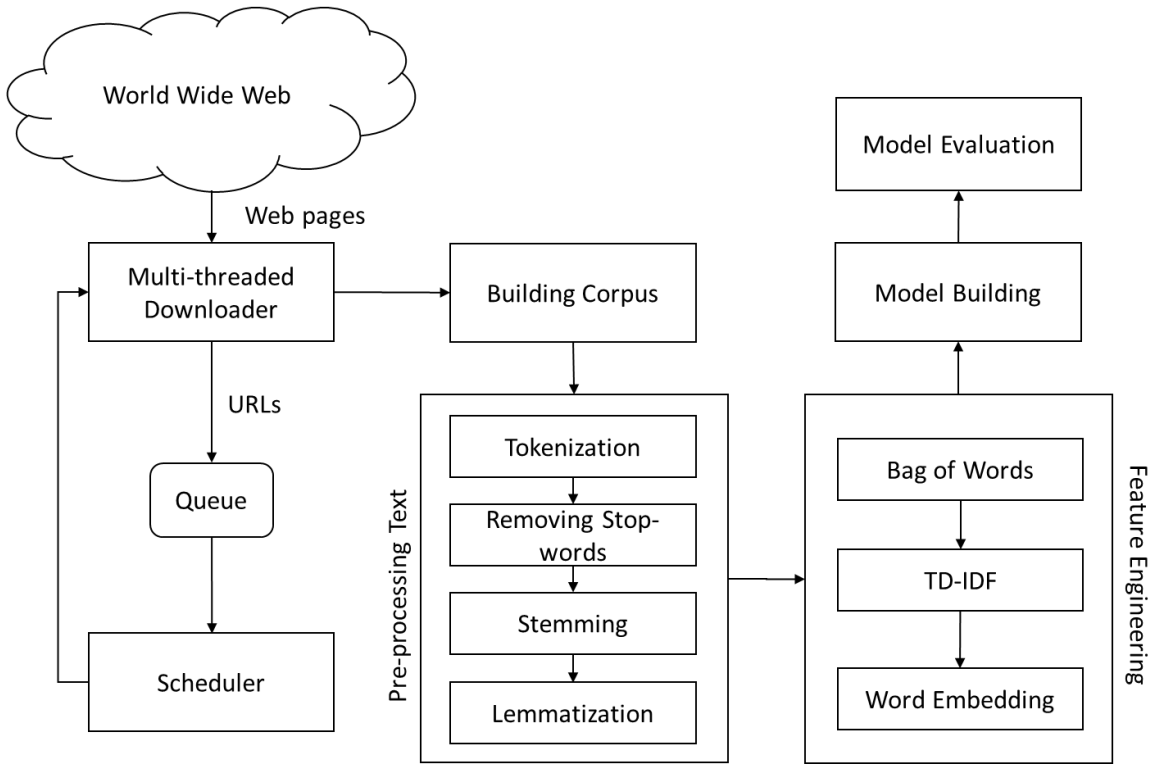

Figure 1. High-level architecture of the proposed model

from online newspapers. The architecture considers the creation of a web crawler, which systematically browses the web recovering for news related to criminal events, based on queries that consider geographic, temporal and media limitations, in order to have a robust body of news to generate intelligence about of the nature of these events.

After the construction of the body of news to analyze, there are the pre-processing text and features engineering phases, the first is disaggregated into the tasks of tokenization, removal of stop-words, stemming and lemmatization, while the second phase is made up of the bag of words, TD-IDF and word embedding tasks. Finally, model building deals with the application of a classification algorithm, after identifying indicators such as commune or nature of the criminal event, and the model evaluation measures the efficiency of the model generated by identifying criminal events within the body of a news.

Web Crawling

Web content has increasingly become a focus for academic research and computer programs are necessary to perform any large-scale processing of web pages, which requires the use of a web crawler to retrieve the websites to analyze (Thelwall, 2001). The web crawler plays a critical role in making the web easier to use. A web crawler is a program that browses the web on behalf of the search engine and downloads web pages for further processing by the search engine. The program receives an initial set of URLs, the pages of which must be downloaded from the web, the crawlers extract the URLs that appear on the retrieved pages and give this information to the crawler control module, which determines which links to visit next and feeds the links to visit the trackers again. Crawlers continue to visit the web until local resources, such as storage, are exhausted, or until user-defined criteria, such as depth level, are met (Mukhopadhyay, 2019).

Summarizing, the procedure developed by a web crawler is to retrieve web addresses from an initial address, analyze the content of the web pages and look for links to new pages, downloading the content of these and repeating the process successively (Achsan and Wibowo, 2014; Alhajj and Rokne, 2018). The developed tool for crawl the web is configured to extract news only from a list of newspapers established by the authors (formal media), to avoid false news or uncredited websites. For the generation of the corpus, the Python's libraries "Beautiful Soup" (Richardson, 2019) and "requests" (Python Software Foundation, 2019) was used. Through the search algorithm, the web was tracked to obtain those news items related to criminal event, bounded to the geographical area of the metropolitan region, specifically near to Santiago city (Chile) and other surrounding communes.

Pre-Processing Text

The language of the collection of linguistic material from press texts is Spanish. The structure of the corpus considered as the domain of the experiment correspond to 5492 news of the main newspapers that cover the events of the sampled area.

In natural language processing (NLP), tokenization is the identification of words or phrases in a text. The tokenizer divides a given article into a list of sentences. Tokenization is a step which splits longer strings of text into smaller pieces, or tokens. Larger chains of text can be tokenized into sentences, sentences can be tokenized into words, etc. Further processing is generally performed after a piece of text has been appropriately tokenized. Tokenization is also referred to as text segmentation or lexical analysis (Verma and Gaur, 2014). After the conversion of the text into a vector of words and/or sentences, punctuation marks and other characters should be removed. In a second derivative, all those words that lack a meaning by themselves or "stop words" are eliminated, stop words are those words which are filtered out before further processing of text, since these words contribute little to overall meaning, given that they are generally the most common words in a language, these empty words are usually articles, prepositions, conjunctions, pronouns, etc.

Additionally, the stemming and lemmatization phase, can be defined as the process that reduces a set of words to its "stem" or common lexical root (Armando et al., 2011). Stemming is the process of eliminating affixes (suffixed, prefixes, infixes, circumfixes) from a word in order to obtain a word stem, while lemmatization is related to stemming, differing in that lemmatization is able to capture canonical forms based on a word's lemma (Kao and Poteet, 2007).

Feature Engineering

Feature engineering is about creating new input features from existing ones, it is possible think of data cleaning as a process of subtraction and feature engineering as a process of addition. While statistical language models, in its essence, are the type of models that assign probabilities to the sequences of words. After pre-processing, it is generated random sequences of n elements, to identify the appearance or frequency of a set of words, by that notion, a 2-gram (or bigram) is a two-word sequence of words like "Quinta Normal", "Estación Central" (by its name in spanish), or "your home", and a 3-gram (or trigram) is a three-word sequence of words. An n-Grams allows to efficiently understand large amounts of text, indicating the frequency with which each sequence of words occurs (Daniel and Martin, 2018). After the construction of n-Grams, it is developed the label allocation phase evaluates whether the news contains certain characteristics and the position in which they are found. The characteristics to evaluate are:

- Does the news have a crime term?
- Do you have a location term?
- Do you have terms of car theft?
- Among others.

These characteristics will be checked in each sentence of the news, and intersect at the end, in order to obtain the set of characteristics that will allow to continue with the analysis of the theft locations (Mawby, 2015) and the analysis of the nature of the criminal event (Moghaddam et al., 2013).

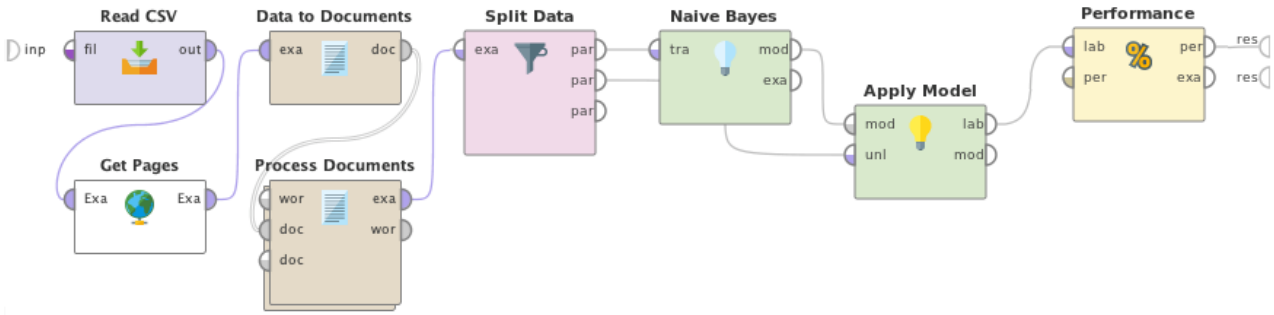

Figure 2. High-level model of news processing from online newspapers

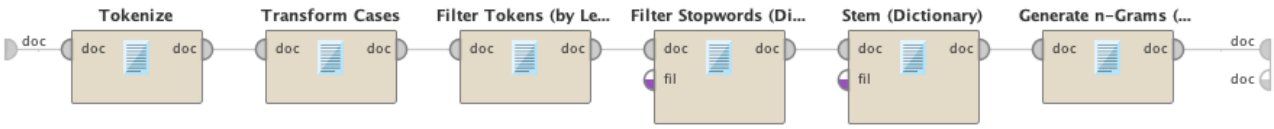

Figure 3. Document processing module (news)

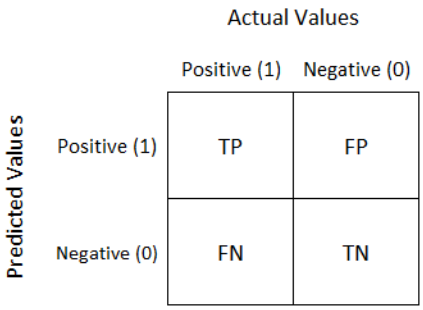

Figure 4. Confusion matrix

Summarizing, the entities extraction is the main step to transform unstructured data into the structured format. An entity named is usually the name of a person, a place, an organization, or an event. The extraction of named entities implies the identification of small pieces of texts and the classification of these into one predefined category. To perform a successful extraction of the entity, a certain amount of preprocessing must be performed on the unstructured data and for each item is necessary identify variables such as location or the nature of the event.

Model Building

After of the building corpus using an algorithm developed in Python (for crawl the web), a model is developed in RAPIDMINER (Rapidminer.com, 2018) to process the news recovered from the web. The high-level model for news processing is presented in **Figure 2**, while the sub-process responsible for the processing of each news is shown in **Figure 3**.

A naïve bayes classifier was used for classifying the commune and the types of events (like a violent crime, vehicle theft or fatalities). Naive Bayes models are a special class of machine learning classification algorithms that are based on a statistical classification technique called the “bayes theorem”. These models are called “naïve” algorithms and assume that the predictor variables are independent of each other, that is, assumes that the presence (or absence) of a feature of a class is unrelated to the presence (or absence) of any other feature. Naive Bayes classifiers can handle an arbitrary number of independent variables, whether continuous or categorical (Darwiche, 2009; Grosan and Abraham, 2011).

Model Evaluation

The analysis of the perception of criminality through the crawl of websites could be an interesting tool for order forces and those responsible for developing public and security policies. For the evaluation of the effectiveness of the proposed model, the statistics Accuracy, Precision, Recall, F1 score and ROC/AUC are considered. Before defining the accuracy precision and recall indicators is necessary introduce the confusion matrix (see **Figure 4**), which is defined as a table that is used to describe the performance of a classification model in a test data set for which the true values are known.

Accuracy (Equation 1) is defined as the fraction of predictions that the model correctly predicted, is the percentage of positive cases detected, Precision (Equation 2) is the ratio of correctly predicted positive observations to the total predicted positive observations, Recall (Equation 3) is the ratio of correctly predicted positive observations to the all observations in actual class - yes, and finally, F1 Score (Equation 4) is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account (Kelleher et al., 2015).

Table 1. Summary of events by geographical zone

Commune	News	Violent crimes	Vehicle Theft	Theft in habited place	Fatalities
Santiago	1109	441	230	89	50
Cerrillos	109	38	20	4	4
Cerro Navia	93	52	10	4	2
Conchalí	99	61	18	5	3
El Bosque	140	80	23	12	3
Estación Central	306	125	42	29	6
Independencia	99	45	26	5	4
La Cisterna	137	68	33	8	5
La Florida	457	143	98	37	18
La Pintana	137	81	20	12	4
La Reina	125	26	41	6	2
Las Condes	378	69	126	20	12
Lo Barnechea	76	16	18	3	1
Lo Espejo	70	44	10	3	1
Macul	116	43	33	11	5
Maipú	350	146	83	19	11
Ñuñoa	251	71	91	9	7
Providencia	421	99	98	32	7
Pudahuel	148	76	27	12	3
Quilicura	192	87	42	7	5
Quinta Normal	130	60	30	6	3
Recoleta	217	132	37	14	9
San Joaquín	72	34	16	5	1
San Miguel	147	58	53	7	5
Vitacura	113	25	50	2	2
Total	5492	2120	1275	361	173

Additionally, sensitivity and specificity are statistical measures of the performance of classification tests, sensitivity measures the proportion of real positives that are correctly identified, while specificity measures the proportion of real negatives correctly identified. The ROC curve, on the other hand, is a graphical representation of sensitivity versus specificity in a binary classifier, where the greater the area under the curve (AUC), the greater the efficiency or predictive capacity of the system (Majnik and Bosnić, 2013).

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score}=\frac{2\times(\text{Recall}\times\text{Precision})}{(\text{Recall}+\text{Precision})} \quad (4)$$

RESULTS AND DISCUSSIONS

Summary of Events

A database of location and nature entities of criminal events was generated according to the set of news samples recovered through the proposed model, identifying the location of certain criminal events by analyzing the semantics of the sentences into news corpus. The summary of the identified events is shown in **Table 1**.

According to the Citizen Security Studies Center of the University of Chile and according to police and penitentiary system statistics, both the number of complaints and detainees show sustained growth in crime (Goldstein Braunfeld ¸ 2003), which has a profound impact on the rates of victimization and fear of the population (Dammert and Lunecke, 2002; Olavarría, 2006). While recent studies, such as the 2019 National Urban Survey of Citizen Security (Instituto Nacional de Estadísticas ¸ 2020) indicate that the victimization rate (Robbery with violence and intimidation, robbery by surprise, robbery with force in the home, theft, injuries, theft or theft of vehicles, theft or theft from vehicles) of homes for crimes of greater social connotation decreased 2.0 percentage points (pp.) compared to 2018, standing at 23.3%, however, the same survey also indicates that 82.0% of the people declared perceiving an increase in crime in the country, that is, 5.2 pp. more than in 2018. Additionally, a report by the Paz Ciudadana foundation indicates that both in Santiago and in regions, the trend is towards increased victimization and fear in the last year (Fundación Paz Ciudadana, 2020).

On the other hand, analyzing the news that was retrieved from the web, tt should be noted that the statistics resulting from the analysis of the news body could have a media bias, because the distribution of criminal events consider only those events that

Table 2. Classification results

	Accuracy	Precision	Recall	F1 Score	AUC/ROC
Violent crimes	0.89286	0.97143	0.87179	0.91892	0.92158
Vehicle Theft	0.90244	0.92063	0.89231	0.90625	0.91170
Theft in habited place	0.86409	0.92455	0.85449	0.88814	0.87352
Fatalities	0.94937	0.92593	0.92593	0.92593	0.90864

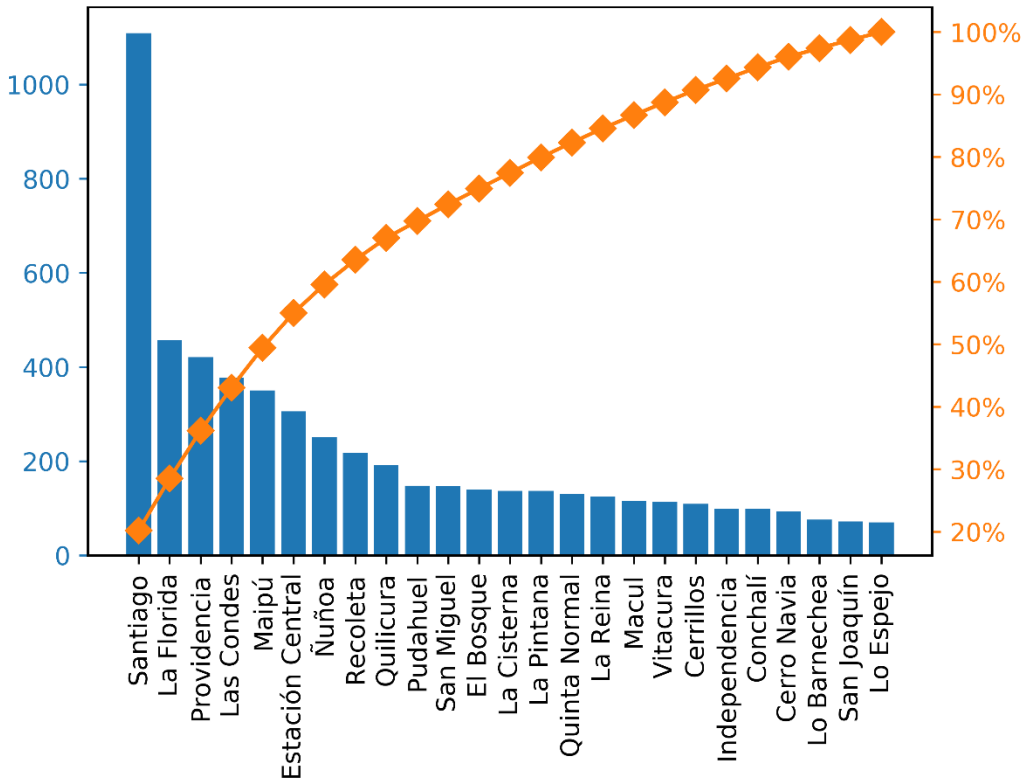

Figure 5. News distribution by commune of the metropolitan region, Chile

generated enough impact. However, the work developed has the potential to identify the perception in the community of the distribution of criminal acts, due to the impact of the information shown through a mass media such as online newspapers.

Table 2 summarizes the statistics that evaluate the performance of the proposed model, measuring the accuracy, precision, recall and predictive capacity of it to identify events categorized as violent crimes, vehicle thefts, theft in habited place and fatalities (resulting in death), demonstrating the ability of the model to recognize criminal events within the body of the news. Statistics indicate that the model is generally efficient in categorizing the nature of criminal events, and the statistics in **Table 2** validate it.

Analysis of the Type of Events

The criminal events considered in this study were the amount and the nature of news covered by the media (newspapers) of the main communes of the metropolitan region, Chile, disaggregating the effects according to vehicle thefts, violent crimes, theft in habited places and/or fatalities, giving as balance that the communes in the ones that occur the most criminal events are “Santiago”, “La Florida”, “Providencia”, “Las Condes”, “Maipú” and “Estación Central” as shown in the Pareto chart of **Figure 5**.

Additionally, **Figure 6** shows the percentage distribution of the types of events, where it is appreciated that approximately 28% of the events correspond to violent crimes (see **Figure 6a**), a 18.8% correspond to vehicle theft (see **Figure 6b**), 22% corresponds to crimes associated with vehicle theft (see **Figure 6c**) and 8% have results of fatality (see **Figure 6d**).

Finally, the heatmaps of criminal events (see **Figure 7**) reported by online newspapers indicate that most criminal events occur in the northwestern areas of “Gran Santiago”. However, it stands out that the set of sampled data is limited and due to the nature of the sources, the results could be mediated bias.

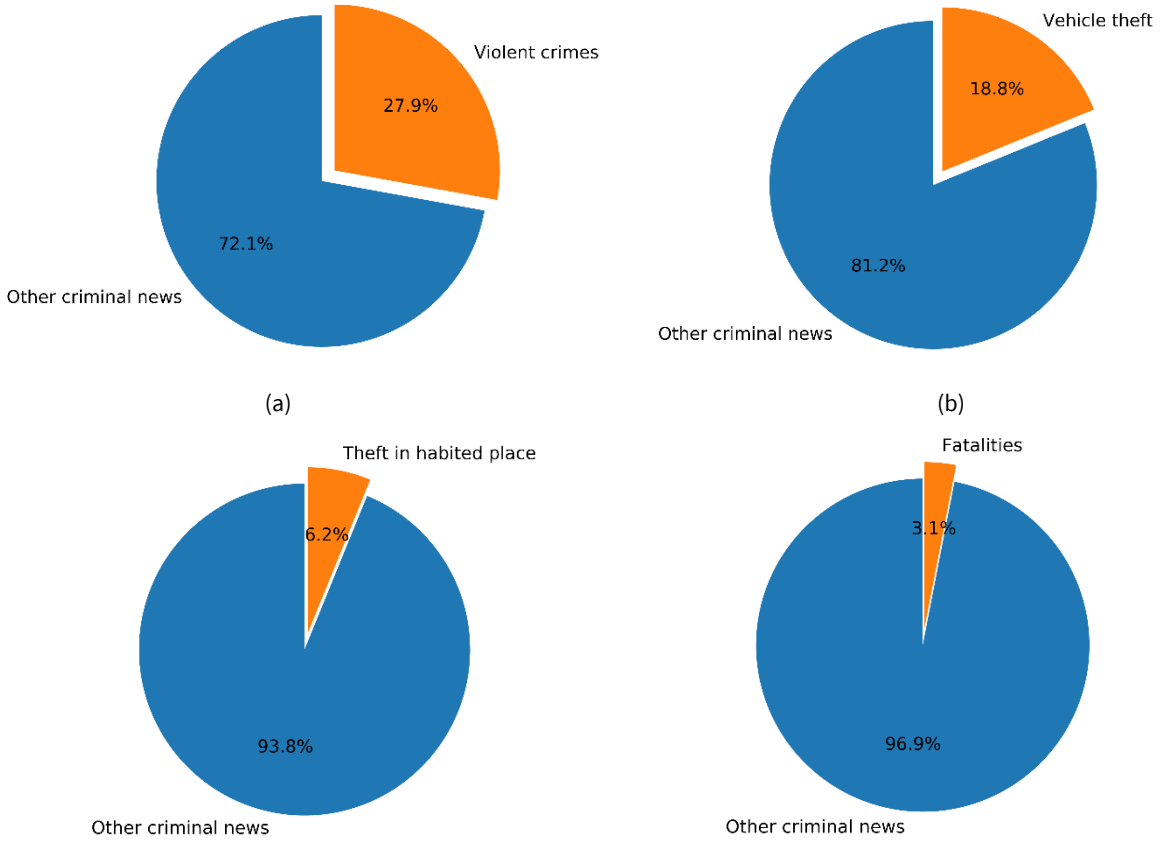

Figure 6. Percentage distribution of events violent crimes (a), vehicle theft (b), theft in habited place (c) and fatalities (d) with respect to the total news of criminal events recovered

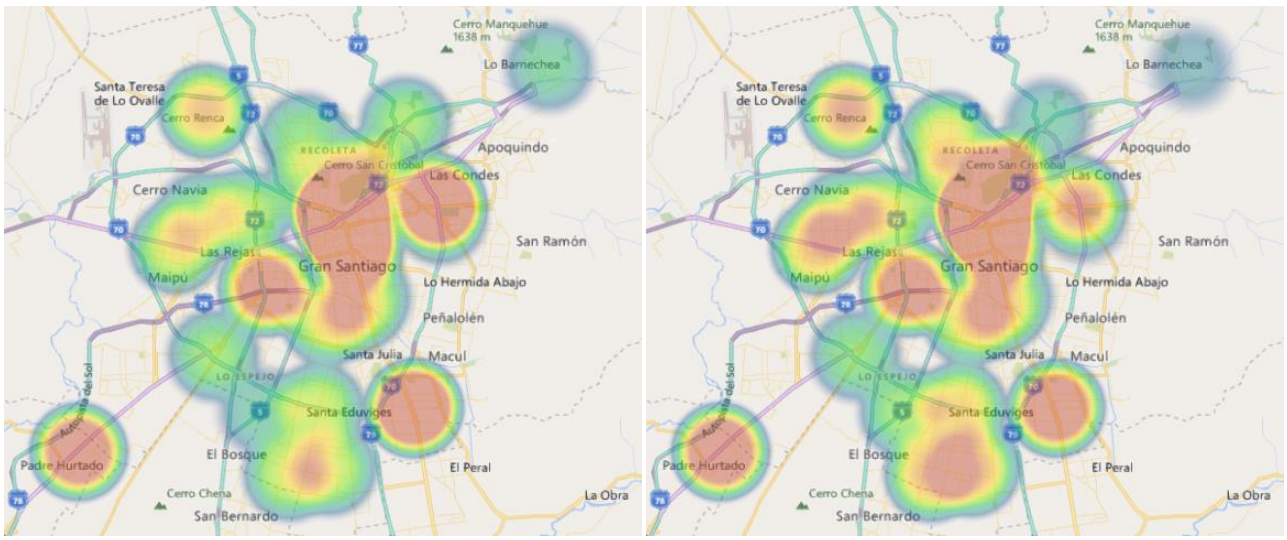

Figure 7. Heatmap of the total criminal events distribution (a), violent crimes (b), vehicle theft (c), theft in habited place (d) and crimes resulting in fatality (e)

Figure 7 (continued). Heatmap of the total criminal events distribution (a), violent crimes (b), vehicle theft (c), theft in habited place (d) and crimes resulting in fatality (e)

CONCLUSIONS

This work presents a simple methodology for the location of the geographical zone in which a criminal event was developed inside the body of the news, in addition to the nature of the event. This models has the potential to carry out crime analysis operations, crime patterns and visualization of the concentration of criminal events, thus contributing to having detailed intelligence on how the perception of crime and fear is distributed among the population, and may become a useful tool for the security forces or those in charge of developing public security policies, so it is expected that the techniques of crime prediction will be implemented in the future to improve the functionality of the crime monitoring system.

The evaluation of the proposed architecture to estimate the perception of crime in the domain of the proposed communes indicates that it is efficient in categorizing the news and the nature of these (validated by the performance indicators), the model reveals the areas that present a greater number of criminal events and the distribution of the types of events with respect to the total number of events retrieved from online newspaper articles. The findings also include a higher incidence of events in the northwest area of ”Gran Santiago”.

The amplitude of the corpus can be improved by expanding the news article crawler, incorporating other sources how blogs or additional web pages that broadcast news. In addition, the task of recognizing and extracting entities can be improved by incorporating more rules that will improve the accuracy and completeness of the entity’s extraction process. Other interesting contributions could be the inclusion or crossing with data retrieved from social networks or the crossing of these with data handled by order and social security agencies.

In order to continue this line of research with the analysis of news from online newspapers, it is planned to incorporate a greater number of sources in the feeding (such as social networks, how twitter or YouTube), the disaggregation of events by limited

geographical areas (streets and/or sectors of a particular commune) and the incorporation of techniques such as sentiment analysis to study in more detail the levels of violence associated to certain criminal events, cataloging documents according to the positive or negative connotation of the language used in it.

REFERENCES

Abdul Jalil, M. @ M., Ling, C. P., Mohamad Noor, N. M. and Mohd., F. (2017). Knowledge Representation Model for Crime Analysis. *Procedia Comput. Sci., 116*, 484-491. https://doi.org/10.1016/j.procs.2017.10.067

Achsan, H. T. Y. and Wibowo, W. C. (2014). A fast distributed focused-web crawling. *Procedia Eng., 69*, 492-499. https://doi.org/10.1016/j.proeng.2014.03.017

Adnan, M., Nagi, M., Kianmehr, K., Tahboub, R., Ridley, M. and Rokne, J. (2011). Promoting where, when and what? An analysis of web logs by integrating data mining and social network techniques to guide ecommerce business promotions. *Soc. Netw. Anal. Min., 1*(3), 173-185. https://doi.org/10.1007/s13278-010-0015-3

Alami, S. and Elbeqqali, O. (2015). Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts. *2015 10th Int. Conf. Intell. Syst. Theor. Appl. SITA 2015*. https://doi.org/10.1109/SITA.2015.7358435

Alhajj, R. and Rokne, J. (2018). *Encyclopedia of Social Network Analysis and Mining*. New York, NY: Springer New York.

Andrienko, G., Gunopulos, D., Ioannidis, Y., Kalogeraki, V., Katakis, I., Morik, K. and Verscheure, O. (2016). Mining Urban Data (Part B). *Inf. Syst., 57*, 75-76. https://doi.org/10.1016/j.is.2016.01.001

Armando, H., Eraso, O., Alberto, C. and Lozada, C. (2011). Stemming in the Spanish Language for Documents Recovered from the Web. (58), 107-114.

Arulanandam, R., Savarimuthu, B. T. R. and Purvis, M. A. (2014). Extracting Crime Information from Online Newspaper Articles. In *Proceedings of the Second Australasian Web Conference (AWC 2014)*, Auckland, New Zealand, Awc, 31-38.

Badal-Valero, E., Alvarez-Jareño, J. A. and Pavía, J. M. (2018). Combining Benford's Law and machine learning to detect money laundering. An actual Spanish court case. *Forensic Sci. Int., 282*, 24–34. https://doi.org/10.1016/j.forsciint.2017.11.008

Belur, J. and Johnson, S. (2018). Is crime analysis at the heart of policing practice? A case study. *Polic. Soc., 28*(7), 768-786. https://doi.org/10.1080/10439463.2016.1262364

Boba, R. (2001). *Introductory Guide to Crime Analysis and Mapping*. Available at: http://www.ncjrs.gov/App/abstractdb/AbstractDBDetails.aspx?id=194685

Boba, R. (2016). *Crime Analysis with Crime Mapping* (4th Ed.). London, UK: SAGE Publications.

Cesur, R., Ceyhan, E. B., Kermen, A. and Sağiroğlu, Ş. (2017). Determination of potential criminals in social network. *Gazi Univ. J. Sci., 30*(1), 121–131.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y. and Chau, M. (2004). Crime data mining: A general framework and some examples. *Computer (Long. Beach. Calif)., 37*(4), 50-56. https://doi.org/10.1109/MC.2004.1297301

Çinar, M. S., Genç, B. and Sever, H. (2019). Identifying criminal organizations from their social network structures. Turkish J. Electr. Eng. Comput. Sci., 27(1), 421–436. https://doi.org/10.3906/elk-1806-52

Dammert, L. and Lunecke, A. (2002). Victimización y Temor en Chile: Revisión Teórico Empírica en Doce Comunas del País. Santiago, Chile. Available at: https://www.cesc.uchile.cl/publicaciones/se_01_victimizacion.pdf

Daniel, J. and Martin, J. H. (2018). N-gram Language Models. In *Speech and Language Processing* (3rd Ed.), pp. 28.

Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks* (1st Ed.). Los Angeles, USA: Cambridge University Press.

De Farias, A. M. G., Cintra, M. E., Felix, A. C. and Cavalcante, D. L. (2018). Definition of Strategies for Crime Prevention and Combat Using Fuzzy Clustering and Formal Concept Analysis. *Int. J. Uncertainty, Fuzziness Knowlege-Based Syst., 26*(3), 429–452. https://doi.org/10.1142/S0218488518500216

de Mendonça, R. R., de Brito, D. F., de Franco Rosa, F., dos Reis, J. C. and Bonacin, R. (2020). A framework for detecting intentions of criminal acts in social media: A case study on twitter. *Inf., 11*(3), 1-40. https://doi.org/10.3390/info11030154

De Sousa Netto, M. C., Pinto, A. L. and Semeler, A. R. (2019). Man and machines against crime: An approach based on visual learning. *Educ. Inf., 35*(3), 251–262. https://doi.org/10.3233/EFI-190280

Dowerah Baruah, T. (2012). Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study. *Int. J. Sci. Res. Publ., 2*(5), 1-10.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv., 4*(1), 1–6. https://doi.org/10.1126/sciadv.aao5580

European Commission (2011). *Crime and deviance in the EU*. Brussels, Belgium. Available at: https://ec.europa.eu/research/social-sciences/pdf/policy_reviews/crime-and-deviance_en.pdf

Evans, J. M. and Kebbell, M. R. (2012). The effective analyst: A study of what makes an effective crime and intelligence analyst. *Polic. Soc., 22*(2), 204–219. https://doi.org/10.1080/10439463.2011.605130

Fundación Paz Ciudadana (2020). Índice Paz Ciudadana. Santiago, Chile. Available at: https://pazciudadana.cl/wp-content/uploads/2019/10/IPC-2019-Conferencia.pdf

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decis. Support Syst., 61*, 115-125. https://doi.org/10.1016/j.dss.2014.02.003

Goldstein Braunfeld, E. (2003). Los Robos con Violencia en el Gran Santiago: Magnitudes y Características. Santiago, Chile. Available at: https://www.cesc.uchile.cl/publicaciones/se_05_goldstein.pdf

Grosan, C. and Abraham, A. (2011). *Intelligent Systems* (1st Ed., vol. 17). Berlin, Heidelberg, Germany: Springer Berlin Heidelberg.

Hammouda, K. M. and Kamel, M. S. (2004) Efficient phrase-based document indexing for web document clustering. *IEEE Trans. Knowl. Data Eng., 16*(10), 1279-1296. https://doi.org/10.1109/TKDE.2004.58

Hotho, A., Nürnberger, A. and Paaß, G. (2005). A Brief Survey of Text Mining. *Ldv Forum, 20*(1), 19-62.

Instituto Nacional de Estadísticas (2020). Encuesta Nacional Urbana de Seguridad Ciudadana 2019. Santiago, Chile. Available at: http://www.seguridadpublica.gov.cl/enusc_2012.html

Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., et al. (2015). Crime analytics: Analysis of crimes through newspaper articles. *MERCon 2015 - Moratuwa Eng. Res. Conf.*, pp. 277-282. https://doi.org/10.1109/MERCon.2015.7112359

Justicia de la Torre, C., Sánchez, D., Blanco, I. and Martín-Bautista, M. J. (2018). Text Mining: Techniques, Applications, and Challenges. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., 26*(04), 553-582. https://doi.org/10.1142/S0218488518500265

Kahya-Özyirmidokuz, E. (2016). Analyzing unstructured Facebook social network data through web text mining: A study of online shopping firms in Turkey. *Inf. Dev., 32*(1), 70-80. https://doi.org/10.1177/0266666914528523

Kao, A. and Poteet, S. R. (2007). *Natural Language Processing and Text Mining* (1st Ed.). London, UK: Springer London.

Kaushik, A. and Naithani, S. (2016). A Comprehensive Study of Text Mining Approach. *IJCSNS Int. J. Comput. Sci. Netw. Secur., 16*(2), 69.

Kelleher, J., Mac Namee, B. and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics* (1st Ed.). London, UK: The MIT Press.

Kumar, S. N. (2015). World towards Advance Web Mining: A Review. *Am. J. Syst. Softw., 3*(2), 44-61. https://doi.org/10.12691/AJSS-3-2-3

Li, Y. and Zhong, N. (2004). Web mining model and its applications for information gathering. *Knowledge-Based Syst., 17*(5-6), 207-217. https://doi.org/10.1016/j.knosys.2004.05.002

Li, Y. S. and Qi, M. L. (2019). An approach for understanding offender modus operandi to detect serial robbery crimes. *J. Comput. Sci., 36*, 101024. https://doi.org/10.1016/j.jocs.2019.101024

Lim, C., Kim, K. J. and Maglio, P. P. (2018). Smart cities with big data: Reference models, challenges, and considerations. *Cities, 82*(February), 86–99, 2018, https://doi.org/10.1016/j.cities.2018.04.011

Lim, M., Abdullah, A., Jhanjhi, N. Z. and Khurram Khan, M. (2020). Situation-Aware Deep Reinforcement Learning Link Prediction Model for Evolving Criminal Networks. *IEEE Access, 8*, 16550–16559. https://doi.org/10.1109/ACCESS.2019.2961805

Lim, M., Abdullah, A., Jhanjhi, N. Z. and Supramaniam, M. (2019a). Hidden link prediction in criminal networks using the deep reinforcement learning technique. *Computers, 8*(1), 1–13. https://doi.org/10.3390/computers8010008

Lim, M., Abdullah, A., Jhanjhi, N., Khurram Khan, M. and Supramaniam, M. (2019b). Link prediction in time-evolving criminal network with deep reinforcement learning technique. *IEEE Access, 7*, 184797–184807. https://doi.org/10.1109/ACCESS.2019.2958873

Lin, Y. L., Yen, M. F. and Yu, L. C. (2018). Grid-based crime prediction using geographical features. *ISPRS Int. J. Geo-Information, 7*(8), 298. https://doi.org/10.3390/ijgi7080298

Liu, B. (2011). *Web Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Majnik, M. and Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intell. Data Anal., 17*(3), 531-558. https://doi.org/10.3233/IDA-130592

Makki, S., Assaghir, Z., Taher, Y., Haque, R., et al. (2019). An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access, 7*, 93010–93022. https://doi.org/10.1109/ACCESS.2019.2927266

Marinescu, V. and Balica, E. (2018). *Migration and Crime Realities and Media Representations*. Springer Nature Switzerland AG.

Mawby, R. I. (2004). Crime and Disorder: Perceptions of Business People in Cornwall, England. *Int. Rev. Vict., 11*(2-3), 313-332. https://doi.org/10.1177/026975800401100207

Mawby, R. I. (2015). Exploring the relationship between crime and place in the countryside *. *J. Rural Stud., 39*, 262-270. https://doi.org/10.1016/j.jrurstud.2014.12.003

Meijer, A. and Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. *Int. J. Public Adm., 42*(12), 1031-1039. https://doi.org/10.1080/01900692.2019.1575664

Moghaddam, A. S., Hosseinkhani, J., Chuprat, S., Taherdoost, H. and Baravati, H. B. (2013). Proposing a framework for exploration of crime data using web structure and content mining. *Res. J. Appl. Sci. Eng. Technol., 6*(19), 3617-3624. http://doi.org/10.19026/rjaset.6.3568

Mukhopadhyay, D. (2019). *Web Searching and Mining* (1st Ed.). Singapore: Springer Singapore.

Nokhbeh Zaeem, R., Manoharan, M., Yang, Y. and Barber, K. S. (2017). Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Comput. Secur., 65*, 50–63. https://doi.org/10.1016/j.cose.2016.11.002

O'Shea, T. C. and Nicholls, K. (2002). *Crime Analysis in America*. Massachusetts, USA. Available at: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/812146-commercialmdhd-truckfuelefficiencytechstudy-v2.pdf

Olavarría, M. (2006). *El crimen en Chile Una mirada desde las víctimas*. Santiago, Chile. Available at: http://www.cesc.uchile.cl/publicaciones/se_13_victimizacion.pdf

Pan, Y., Tian, Y., Liu, X., Gu, D. and Hua, G. (2016). Urban Big Data and the Development of City Intelligence. *Engineering, 2*(2), 171–178. https://doi.org/10.1016/J.ENG.2016.02.003

Petherick, W. (2015). *Applied Crime Analysis* (1st Ed.). Waltham, Massachusets, USA: Elsevier.

Pinto, P., Theodoro, I., Arrais, M. and Oliveira, J. (2017). Data mining and social web semantics: A case study on the use of hashtags and memes in Online Social Networks. *IEEE Lat. Am. Trans., 15*(12), 2276-2281. https://doi.org/10.1109/TLA.2017.8071088

Po, L. and Rollo, F. (2018). Building an Urban Theft Map by Analyzing Newspaper Crime Reports. In *Proceedings - 13th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2018*, pp. 13–18. https://doi.org/10.1109/SMAP.2018.8501866

Python Software Foundation (2019). "Requests: HTTP for HumansTM". Available at: https://pypi.org/project/requests/ (Accessed: 11 July 2019).

Rapidminer.com (2018). Available at: https://rapidminer.com/products/ (Accessed: 27 June 2018).

Richardson, L. (2019). *Beautiful Soup: We called him Tortoise because he taught us*. Available at: https://www.crummy.com/software/BeautifulSoup/ (Accessed: 4 March 2020).

Rouvroy, A. (2016). Of Data and Men' Fundamental Rights and Fredoms in a World of Data. Strasbourg, Germany. Available at: https://rm.coe.int/16806a6020

Saldaña, M., Escobar, C., Galvez, E., Torres, D. and Toro, N. (2020). Mapping of the Perception of Theft Crimes from Analysis of Newspaper Articles Online. In *15th Iberian Conference on Information Systems and Technologies (CISTI)*, June, pp. 1-7. https://doi.org/10.23919/CISTI49556.2020.9141154

Saldaña, M., Flores, V., Toro, N. and Leiva, C. (2019). Representation for a prototype of recommendation system of operation mode in copper mining. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, June, 1-4. https://doi.org/10.23919/CISTI.2019.8760884

Sanders, C. and Condon, C. (2017). Crime analysis and cognitive effects: the practice of policing through flows of data. *Glob. Crime, 18*(3), 237–255. https://doi.org/10.1080/17440572.2017.1323637

Sathyadevan, S., Devan, M. S. and Surya Gangadharan, S. (2014). Crime analysis and prediction using data mining. *1st Int. Conf. Networks Soft Comput. ICNSC 2014 - Proc., August*, 406-412. https://doi.org/10.1109/CNSC.2014.6906719

Sidana, A. and Aggarwal, H. (2017). Review of web usage of data mining in web mining. *Int. J. Adv. Res. Comput. Sci., 8*(5), 1-5.

Song, J., Song, T. M., Seo, D. C. and Jin, J. H. (2016). Data Mining of Web-Based Documents on Social Networking Sites That Included Suicide-Related Words Among Korean Adolescents. *J. Adolesc. Heal., 59*(6), 668-673. https://doi.org/10.1016/j.jadohealth.2016.07.025

Spadon, G., Scabora, L. C., Oliveira, P. H., Araujo, M. V. S., et al. (2017). Behavioral Characterization of Criminality Spread in Cities. *Procedia Comput. Sci., 108*, 2537-2541. https://doi.org/10.1016/J.PROCS.2017.05.118

Stenton, A. E. (2006). *Crime Analysis: An Examination of Crime Prevention and Reduction Strategies*. University of Ottawa.

Talib, R., Kashif, M., Ayesha, S. and Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *Int. J. Adv. Comput. Sci. Appl., 7*(11), 414-418. https://doi.org/10.14569/ijacsa.2016.071153

Thelwall, M. (2001). A web crawler design for data mining. *J. Inf. Sci., 27*(5), 319-325. https://doi.org/10.1177/016555150102700503

Tollenaar, N. and van der Heijden, P. G. M. (2009). Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society, Statistics in Society, Series A, 176*(2), 565-584. https://doi.org/10.1111/j.1467-985X.2012.01056.x

Tollenaar, N. and Van Der Heijden, P. G. M. (2019). Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLOS ONE, 14*(3), e0213245. https://doi.org/10.1371/journal.pone.0213245

Vellani, K. H. (2010). *Crime Analysis: for Problem Solving Security Professionals in 25 Small Steps*. Available at: https://popcenter.asu.edu/sites/default/files/library/reading/PDFs/crimeanalysis25steps.pdf

Verma, T. and Gaur, D. (2014). Tokenization and Filtering Process in RapidMiner. *Int. J. Appl. Inf. Syst., 7*(2), 16-18. https://doi.org/10.5120/ijais14-451139

Vishal Gupta, G. S. L. (2009). A Survey of Text Mining Techniques and Applications. *J. Emerg. Technol. web Intell., 1*(1), 17. https://doi.org/10.4304/jetwi.1.1.60-76

Vishwakarma, R. K. and Shankar, R. (2014). Modeling brain and behavior of a terrorist through fuzzy logic and ontology. *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, 857-861. https://doi.org/10.1109/IEEM.2013.6962533

Yang, X., Sun, N., Zhang, Y. and Kong, D. (2008). General Framework for Text Classification Based on Domain Ontology. *Semant. Media Adapt. Pers. 2008. SMAP '08*. Third Int. Work., pp. 147–152. https://doi.org/10.1109/SMAP.2008.17