Publication status: Preprint has been published in a journal as an article
DOI of the published article: https://doi.org/10.51359/2965-4661.2023.259008

Exploring Text Mining and Analytics for Applications in Public Security: An in-depth dive into a systematic literature review

Victor Diogho Heuer de Carvalho, Ana Paula Cabral Seixas Costa

https://doi.org/10.1590/SciELOPreprints.3518

Submitted on: 2022-01-20
Posted on: 2023-01-19 (version 2)
(YYYY-MM-DD)

Exploring Text Mining and Analytics for Applications in Public Security: an in-depth dive into a systematic literature review

Victor Diogho Heuer de Carvalho [a,b,*], **Ana Paula Cabral Seixas Costa** [b]

[a] Universidade Federal de Alagoas, Campus do Sertão, Rod. AL-145, n. 3849, Cidade Universitária, Delmiro Gouveia – Alagoas, Brazil, Postal Code: 57480-000. E-mail: victor.carvalho@delmiro.ufal.br. https://orcid.org/0000-0003-2369-7317

[b] Universidade Federal de Pernambuco, Department of Management Engineering, Av. da Arquitetura, s/n, Cidade Universitária, Recife – Pernambuco, Brazil, Postal Code: 50740-550. E-mail: apcabral@cdsid.org.br. https://orcid.org/0000-0002-2932-4784

* Corresponding author.

Authorship statement

Victor Diogho Heuer de Carvalho: Conceptualization, Methodology, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing.

Ana Paula Cabral Seixas Costa: Supervision, Resources, Validation, Methodology, Writing - Original Draft, Writing - Review & Editing.

Declaration of interests

The authors have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This article was funded by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES, Brazil) (Finance Code 001), the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq, Brazil), and the *Universidade Federal de Alagoas* (UFAL, Brazil).

Abstract

Text mining and related analytics emerge as a technological approach to support human activities in extracting useful knowledge through texts in several formats. From a managerial point of view, it can help organizations in planning and decision-making processes, providing information that was not previously evident through textual materials produced internally or even externally. In this context, within the public/governmental scope, public security agencies are great beneficiaries of the tools associated with text mining, in several aspects, from applications in the criminal area to the collection of people's opinions and sentiments about the actions taken to promote their welfare. This article reports details of a systematic literature review focused on identifying the main areas of text mining application in public security, the most recurrent technological tools, and future research directions. The searches covered four major article bases (Scopus, Web of Science, IEEE Xplore, and ACM Digital Library), selecting 194 materials published between 2014 and the first half of 2021, among journals, conferences, and book chapters. There were several findings concerning the targets of the literature review, as presented in the results of this article.

Keywords: Text mining. Public security. Systematic literature review. Technologies. Future research directions.

1. Introduction

Text mining and analytics is a trendy field both from the research point of view and as a practical application in organizations, being involved with the discovery of knowledge in textual bases and the subsequent analyses that can be applied to texts within a collection. Any organization can benefit from the associated tools, regardless of whether it is private or public. In the public and governmental context, managers can apply these tools to support planning processes and decision-making, using, for example, information extracted from texts extracted from the social web containing opinions coming directly from service users.

It is here that the context of public security can be inserted, using text mining tools to support investigative processes, detection or prediction of criminal activities, prevention of public events harmful to citizens, or even to collect opinions from these citizens about security actions that are already being applied. According to their core functionality, text mining-related approaches contain techniques and methods that can be categorized into information retrieval and extraction, document clustering and classification, natural language processing, and web mining (Talib et al., 2016; Zainol et al., 2018).

Each approach includes different applications on issues related to public security. For instance, Information retrieval and extraction tasks can be applied in the mining process for digital investigation systems to obtain the most relevant texts or documents according to a specific investigational subject (Gowri et al., 2014). Another example is selecting unstructured criminal judgment texts to extract information necessary to support lawyers and judges in their activities (Iftikhar et al., 2019).

The clustering and classification of texts consist of applying techniques capable of creating groups or applying labels to texts according to similar classes or categories (Park et al., 2019). Both tasks were applied by Kuang et al. (2017) in their text mining application to discover meaningful latent crime classes using textual records of crime events. In another example, Das and Das (2019) applied a graph-based clustering technique to discover crime reports' labels based on extracted paraphrasing from large untagged crime corpora. Thao et al. (2017) applied an approach to classifying landing and distribution domains, which were essential components for analyzing cyberattack types.

Natural language processing is a vast area dedicated to analyzing and synthesizing spoken and written language using computers (Jackson & Moulinier, 2002). Another concept that can be given for natural language processing is that it attempts to extract a fuller meaning representation from texts, using linguistic concepts such as part-of-speech and grammatical structure (Kao & Poteet, 2007). It comprises many techniques widely used in text mining to support several other tasks, such as those mentioned above. However, according to Sankar et al. (2020), the essence of natural language processing is identifying the language in a text and applying rules to identify and extract useful information.

In their study, Gravanis et al. (2019) provide an example of fake news detection, proposing a natural language processing approach that uses linguistic-based features combined with machine learning. Their conclusions establish that this type of approach provides publishers with a basis for determining which content should be best evaluated, avoiding publishing untrue information. This note by the authors only reinforces the support that text mining provides when the human analytical capacity fails to detect elements or even events capable of causing damage; in this case, it is also related to more specific issues of cybersecurity involving people using online platforms to communicate and share information.

Web mining is defined as a set of text mining techniques that learn how information is distributed across the Internet through characterization and classification of web pages, enabling discovering how these pages are interrelated (Han et al., 2011). Three types of web mining were reviewed by Kavita et al. (2016) and Kumar and Palanisamy (2008), including content mining, structure mining, and usage mining. These methods were reinforced by de la Torre et al. (2018), who set web mining, or "web data mining," as being equivalent to "text mining on web documents." Still, within this approach, the social web contains a wide range of data to be explored, enabling, for example, the users' sentiment analysis about a given topic (Costa et al., 2012). The works by Anwar and Abulaish (2014a) and Anwar and Abulaish (2014b) contain examples of a web mining application dedicated to retrieving relevant information from webpages for the detection of namesakes, a relevant cybersecurity task supporting the disambiguation of people with the same name.

The literature presented in this introduction exemplifies the wide variety of applications that text mining offers for public security and related matters, supporting a systematic literature review to identify opportunities for new research.

The purpose of this text is to present the details about the results of a systematic literature review on text mining applied to public security, corroborating the validity and importance of associated tools to assist in decisions and actions of public security agencies. The rest of this article is structured as follows: Section 2 details the methodology applied for the systematic literature review; Section 3 presents the details of the review results; Section 4 contains the conclusion of this work.

2. Review Methodological Details

This section provides details about the systematic literature review process, including explanations about: search criteria, filtering, full papers assessment, information extraction, and software used to support analysis. The guidelines applied for the review process were presented by Kitchenham and Charters (2007), following the workflow indicated in Figure 1.

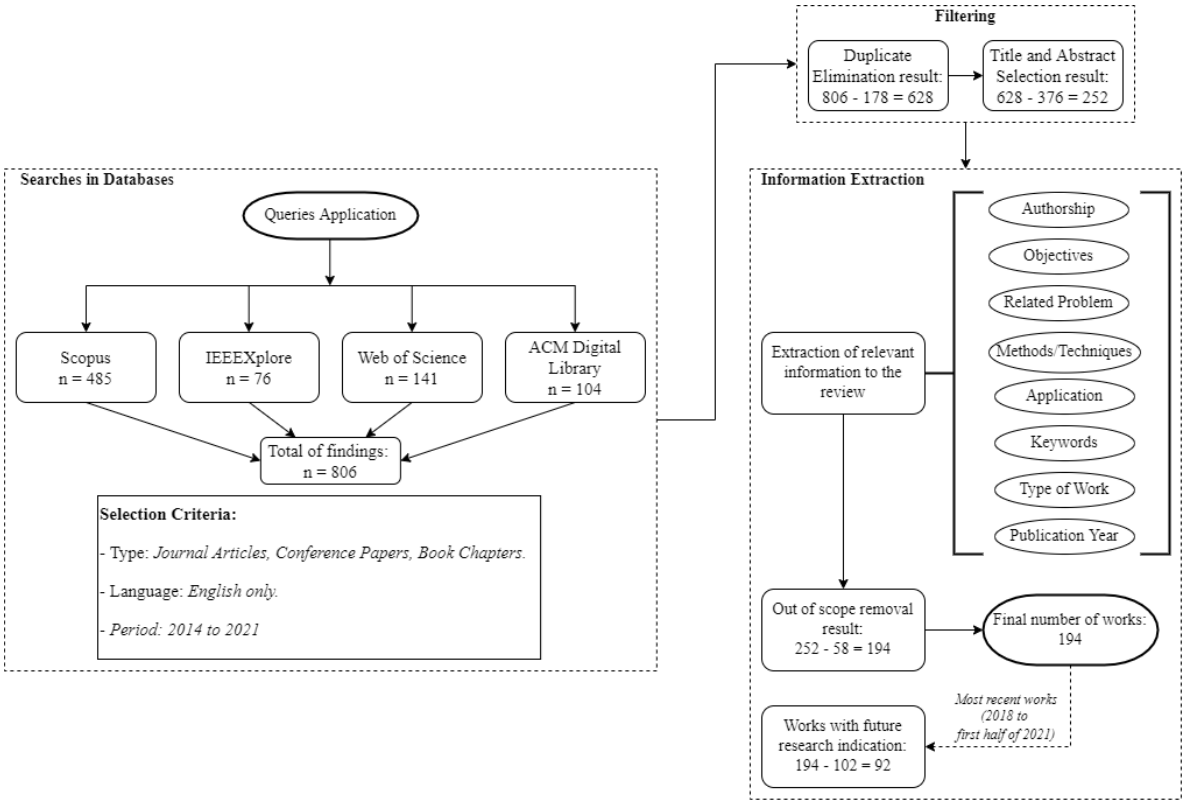

Figure 1. Literature review workflow.

This entire procedure was dedicated to extracting and selecting literature aiding in identifying a) the main areas of application of text mining in public security; b) the most recurrent text mining technologies in public security-related issues, according to selected literature; and c) open opportunities and challenges for the application of text mining in public security, according to future directions pointed out in the most recent literature. The following sections will describe each part of the workflow in Figure 1.

3.1 Literature selection criteria and search terms

The literature search was limited by the following criteria: a) works only in English, b) journal articles, complete conference papers and book chapters, and c) publication dates between 2014 and 2021.

For selection criterion c), the period of eight years was defined by considering two aspects of the literature. The first was to filter the most recent literature published on the themes focused on this work. The second concerns the number of works published on the theme because 70.62% is concentrated between 2017 and 2020. The previous three years (2014, 2015, and 2016) add up to only 27.84% of the works, and the year 2021 completes the total number of selected works (representing only 1.55% of them) since the literature extraction was performed before this year had ended.

The search queries considered the following Boolean condition: *("text mining") AND ("public security" OR "crime" OR "terrorism" OR "piracy" OR "drug trafficking" OR "arms trafficking" OR "human trafficking" OR "sexual exploitation" OR "prostitution" OR "pedophilia" OR "rape" OR "homicide" OR "murder" OR "femicide" OR "infanticide" OR*

"bodily injury" OR "extortion" OR "theft" OR "robbery" OR "assault" OR "burglary" OR "property damage" OR "misappropriation" OR "money laundering" OR "embezzlement" OR "stellionate" OR "receiving" OR "kidnapping" OR "defamation" OR "cybercrime").

The crime-related terms in the query were taken from the Brazilian Penal Code (Decree-Law No. 2848 / 1940), which is still the current penal code in Brazil, updated by a series of other legal provisions over the years and defines all possible penalties for the related crimes.

To corroborate these terms, the American Model Penal Code also provides descriptions of several of these terms so that others are either variations or, even if not included, are related to the present ones, as can be seen in Robinson and Dubber (2007). Examining the crime information related to the Crown Prosecution Service[1] (England and Wales), these terms are also corroborated by the several categories listed: domestic abuse, drug offenses, fraud, and economic crime, hate crime, sexual offenses, terrorism, verbal abuse, violent crime, and even cybercrime. As a final resource to corroborate the presence of these terms, Brazilian Jurisprudence[2] was also consulted through documents of court sentences on crimes.

This condition was applied to four databases, considering the specific formats of each one: Scopus, Web of Science, IEEE Xplore, and ACM Digital Literature.

3.2 Filtering

After retrieving the literature from each database, sequential filters were applied to eliminate duplicates, resulting in the exclusion of 178 works, leaving for the subsequent analysis only those related to the theme through the direct screening of titles and abstracts. The titles and abstracts screening resulted in the exclusion of 376 works for the information extraction phase.

3.3 Information extraction

The information extraction strategy was based on the manual search for objectives, methodology, applications, and other indications about what each selected works applied. A spreadsheet was maintained to include information about authorship, publication channel name, primary objectives and purposes, related problems, synthesis of the methods, techniques, technologies used, key applications, keywords, type of work, publication year, and key terms related to the techniques and technologies. Also, a final field was included to identify proposals for future work if they are described in the publication.

During the assessment of full papers for extracting information, 58 papers were excluded as out of the review's scope: (i) for not adding relevant information to the research since they were literature reviews, although focusing on themes related to text mining in the public security; (ii) because they were in a language other than English; and (iii) because they applied other types of analysis in public security that did not involve text mining.

Thus, this phase included generating a spreadsheet containing summaries of the 194 selected works containing relevant information following the process outlined in the literature selection process. All the information extraction from the selected works was

[1] See in the Crown Prosecution Service website: https://www.cps.gov.uk/cps/crime-info.
[2] Through the Brazilian Jurisprudence portal: https://www.jusbrasil.com.br/

performed manually by reading the texts and recording the desired information in the spreadsheet. Among these 194 works, just 92 entered for the opportunities and challenges extraction since they contained directions for future research.

3.4 Software used and extracted information availability

Elsevier's Mendeley software was leveraged for the screening process and assisted in the insertion of citations creation of references. A Python script was applied to make counting create charts and tables from the spreadsheet with information extracted from the literature. Web application WordItOut[3] was used to obtain the word cloud with the most frequent terms related to the technologies and techniques found in the literature.

The information extracted from the 194 selected works is available in a spreadsheet within a publicly open repository at GitHub[4]. This spreadsheet contains the information of interest for each selected article, as described in Figure 1.

3. Systematic Review Results Detailing

This section contains details about the literature review results, separating them according to the three items defined in the previous section:

a) The main areas of application of text mining in public security.
b) The most recurrent text mining technologies in public security.
c) The open opportunities and challenges for text mining in public security based on future directions in the most recent literature (2018 to the first semester of 2021).

For a complete listing of journals, conferences, and books, with the related counts for each channel, see Tables A1, A2, and A3 in Appendix A.

3.1 Applications areas for text mining in public security

Literature provided an understanding of the main areas related to public security demanding text mining applications. Listing from the area with the greatest number of selected works to the one with fewer works: cybersecurity (62 works), general crime detection/prediction (29), fraud detection (22), terrorism detection (16), cyberbullying detection (14), digital/cyber forensics (14), support to the judiciary power (6), support to law enforcement agencies' actions (6), crimes' victims support (4), sex-related crimes detection (4), drug-related crimes detection (3), espionage detection (3), information security (3), software piracy detection (2), civil unrest detection (2), drug-related crimes detection and weapons' trafficking detection (1), weapons' trafficking detection (1), armed conflicts solution (1), and violence against woman analysis (1). Note that there is one work in an area combining “drug-related crimes detection and weapons' trafficking detection”.

The tables below present the works according to their publication years according to these application areas. Table 1 contains the details of the "Cybersecurity" area.

[3] Accessible through the URL: https://worditout.com/
[4] To access the repository, use the URL: https://github.com/victorheuer/tm_ps_literature-info

Table 1. Selected works in the Cybersecurity area, separated by year of publication.

Cybersecurity		
Year	Works	Count
2014	(Anwar & Abulaish, 2014a; Anwar & Abulaish, 2014b; Li, Xiao, et al., 2014; Suarez-Tangil et al., 2014)	4
2015	(Alami & Elbeqqali, 2015; Chang & Wang, 2015; Choudhary & Vidyarthi, 2015; Hernandez-Castro & Roberts, 2015; Kumar et al., 2015; Sundarkumar et al., 2015; Trovati et al., 2015; Zareapoor & Seeja, 2015)	8
2016	(Agrawal & Kaushal, 2016; Chen et al., 2016a; Chen et al., 2016b; Hao and Dai, 2016; Martin et al., 2016; Tayal & Ravi, 2016)	6
2017	(Alothman & Rattadilok, 2017; Barbon Jr. et al., 2017; Cardoza & Wagh, 2017; Fa et al., 2017; Hadad et al., 2017; Husari et al., 2017; Martinelli et al., 2017; Marulli & Mercaldo, 2017; Samtani et al., 2017; Sharmin & Zaman, 2017; Thao et al., 2017; Vidyarthi et al., 2017)	12
2018	(Ariffin et al., 2018; Chung et al., 2018; Concepción-Sánchez et al., 2018; Dong et al., 2018; Elkhawas & Abdelbaki, 2018; Hadad et al., 2018; Husari et al., 2018; Ruano-Ordás et al., 2018; Samtani et al., 2018; Silomon & Roeling, 2018; Sonowal & Kuppusamy, 2018; Zainal et al., 2018)	12
2019	(Halouzka & Buřita, 2019; Mohasseb et al., 2019; Niekerk et al., 2019; Roopa & Induja, 2019; Sameera & Vishwakarma, 2019; Balim & Gunal, 2019; de Boer et al., 2019; Gravanis et al., 2019; Sudha & Rupa, 2019; Kakavand et al., 2019; Palad et al., 2019)	11
2020	(Alagheband et al., 2020; Al-Ramahi et al., 2020; Battaglia et al., 2020; Calderon et al., 2020; Ma et al., 2020; Mishra et al., 2020; Samtani et al., 2020; Palad et al., 2020; Pires & Georgieva, 2020)	9
	Total	**62**

The Cybersecurity area had the largest number of works selected in 2017 and 2018, with the same amount, followed by 2019. These three years represent 56.45% of the total within this area. Table 2 contains the details in the "General crime detection/prediction" area.

Table 2. Selected works in the General crime detection/prediction area, separated by year of publication.

General crime detection / prediction		
Year	Works	Count
2014	(Alruily et al., 2014; Seidler et al., 2014)	2
2015	(Sharef et al., 2015)	1
2016	(Marivate & Moiloa, 2016; Wajid & Samet, 2016)	2
2017	(Almehmadi et al., 2017; Das & Das, 2017a; Das & Das, 2017b; Das & Das, 2017c; Kuang et al., 2017; Netsuwan & Kesorn, 2017)	6
2018	(Aghababaei & Makrehchi, 2018; AL-Saif & Al-Dossari, 2018; Percy et al., 2018; Po & Rollo, 2018; Ristea et al., 2018; Tran & Tran, 2018)	6
2019	(Das & Das, 2019; Mine et al., 2019; Nedeljkovic et al., 2019; Qazi & Wong, 2019; Savaş & Topaloğlu, 2019; Malim et al., 2019; Kaur et al., 2019)	7
2020	(Alatrista-Salas et al., 2020; Birks et al., 2020; Mukherjee & Sarkar, 2020; Lal et al., 2020; Saldana et al., 2020)	5

	Total	**29**

In the "General crime detection/prediction" area, 2019 concentrated the largest number of selected works, followed by 2017 and 2018, each with the same amount. These three years represent 65.52% of the total within this area. Table 3 demonstrates the works selected for the "Fraud detection" area.

Table 3. Selected works in the Fraud detection area, separated by year of publication.

Fraud detection		
Year	Works	Count
2014	(Li, Xu, et al., 2014; Yang et al., 2014; Zaki & Theodoulidis, 2014)	3
2016	(Dong, Liao, et al., 2016; Dong, Xu, et al., 2016; Saha et al., 2016)	3
2017	(Chen et al., 2017; Fontanarava et al., 2017; Hajek & Henriques, 2017; Owda et al., 2017; Savaliya & Philip, 2017; Zaeem et al., 2017)	6
2018	(Bhardwaj & Gupta, 2018; Lee et al., 2018; Maktabar et al., 2018)	3
2019	(Dastjerdi et al., 2019; Rabuzin & Modrušan, 2019; Sahu et al., 2019)	3
2020	(Angenent et al., 2020; Kabwe & Phiri, 2020; Monish & Pandey, 2020)	3
2021	(Siering et al., 2021)	1
	Total	**22**

In "Fraud detection", 2017 concentrated the largest number of selected works, representing 27.27% of the total in this area. The years 2014, 2016, 2018, 2019, and 2020 had the same amount of work, representing 68.18% within the area. Table 4 contains the details of selected literature from the "Terrorism detection" area.

Table 4. Selected works in the Terrorism detection area, separated by year of publication.

Terrorism detection		
Year	Works	Count
2015	(Johnston & Weiss, 2017; Overbeck, 2015)	2
2017	(Cataldo et al., 2017; Lekea & Karampelas, 2017; Subhan et al., 2017; Tutun et al., 2017)	4
2018	(Al-Khalisy & Jehlol, 2018; Alguliyev et al., 2018; Mansour, 2018; Öztürk & Ayvaz, 2018; Zahra et al., 2018)	5
2019	(Bisgin et al., 2019; Petrovskiy & Chikunov, 2019)	2
2020	(Castillo-Zúñiga et al., 2020; Rehman et al., 2020; Miranda et al., 2020)	3
	Total	**16**

In "Terrorism detection", 2018 had the highest number of works, followed by 2017, totaling 56.25% of all works in this area. Table 5 contains details on the "Cyberbullying detection" area.

Table 5. Selected works in the Cyberbullying detection area, separated by year of publication.

Cyberbullying detection		
Year	Works	Count
2014	(Margono et al., 2014)	1
2016	(Zhao et al., 2016)	1
2017	(Chandra et al., 2017; Noviantho et al., 2017; Zhao & Mao, 2017)	3

2018	(Alakrot et al., 2018; Ventirozos et al., 2018)	2
2019	(Andleeb et al., 2019)	1
2020	(Adikara et al., 2020; Wang et al., 2020; Slamet et al., 2020; Ishara Amali & Jayalal, 2020)	4
2021	(Bozyiğit et al., 2021; Choi et al., 2021)	2
	Total	**14**

In "Cyberbullying detection", 2020 presented the highest number of selected works, followed by 2017, with these two years adding up to 50% of the total in this area. Table 6 separates the selected works within the "Digital/Cyber forensics" area.

Table 6. Selected works in the Digital/Cyber forensics area, separated by year of publication.

Digital / Cyber forensics		
Year	Works	Count
2014	(Gowri et al., 2014; Noel & Peterson, 2014)	2
2015	(Ding et al., 2015; Spitters et al., 2015)	2
2016	(Aboluwarin et al., 2016; Alami & Elbeqqali, 2016; Hicks et al., 2016; Li et al., 2016; Yang et al., 2016)	5
2017	(Venčkauskas et al., 2017; Xylogiannopoulos et al., 2017)	2
2018	(Giacalone et al., 2018)	1
2019	(Tajuddin et al., 2019; Henseler & Hyde, 2019)	2
	Total	**14**

The year 2016 concentrated the largest number of works selected in the "Digital/Cyber forensics", representing 35.71% of the total in this area. The years 2014, 2015, 2017, and 2019 had the same amount of work, adding up to 57.14% of the works in this area. Table 7 concerns the "Support to the Judiciary power" area.

Table 7. Selected works in the Support to the Judiciary power area, separated by year of publication.

Support to the Judiciary power		
Year	Works	Count
2015	(Nikolić et al., 2015)	1
2019	(Iftikhar et al., 2019; Pina-Sánchez et al., 2019a; Pina-Sánchez et al., 2019b; Xia et al., 2019)	4
2020	(Gomes & Ladeira, 2020)	1
	Total	**6**

In "Support to the Judiciary power", the year 2019 concentrated 66.67% of the works in the area. Each of the two other years that appeared, 2015 and 2020, has only one selected work. Table 8 shows the breakdown of the "Support to Law Enforcement agencies actions" area.

Table 8. Selected works in the Support to Law Enforcement agencies actions area, separated by year of publication.

Support to Law Enforcement agencies actions		
Year	Works	Count

2014	(Badii et al., 2014)	1
2015	(Bisio et al., 2015)	1
2019	(Basilio et al., 2019; Behmer et al., 2019)	2
2020	(Basilio et al., 2020; Hou et al., 2020)	2
	Total	**6**

In "Support to Law Enforcement agencies actions", 2019 and 2020 had the same amount of work, representing 66.67% of the total in this area. More than one area will be presented per table from the following table, as the sum of the totals of the areas that appear together does not exceed 10. Table 9 refers to "Crimes' victims support" and "Sex-related crimes ".

Table 9. Selected works in the Crimes' victims support and Sex-related crimes areas, separated by year of publication.

Crimes' victims support		
Year	Works	Count
2018	(Karystianis et al., 2018)	1
2019	(Karystianis et al., 2019)	1
2020	(Lyu et al., 2020; Karystianis et al., 2020)	2
	Total	**4**
Sex-related crimes detection		
2014	(Parapar et al., 2014)	1
2015	(Miah et al., 2015)	1
2018	(Andriansyah et al., 2018, Hultgren et al., 2018)	2
	Total	**4**

The "Crimes' victims support" area had 50% of its works in 2020, while the "Sex-related crimes detection" area concentrated 50% of its works in 2018. Table 10 contains the selected works from the "Drug-related crimes detection", "Espionage detection", and "Information security" areas.

Table 10. Selected works in the Drug-related crimes detection, Espionage detection, and Information security areas, separated by year of publication.

Drug-related crimes detection		
Year	**Works**	**Count**
2016	(Neumann & Sartor, 2016)	1
2017	(Nguyen et al., 2017)	1
2018	(Cichosz, 2018)	1
	Total	**3**
Espionage detection		
2016	(Liang et al., 2016; Liang & Biros, 2016)	2
2020	(Glancy et al., 2020)	1
	Total	**3**
Information security		
2015	(Xianghui et al., 2015)	1
2016	(Lee et al., 2016; Nwafor et al., 2016)	2
	Total	**3**

While in "Drug-related crimes detection" every year had the same amount of work published, in "Espionage detection" and "Information security" the year 2016 had the largest amount of work.

In "Software piracy detection", only occurred the works by Kim et al. (2018) and Sarwar et al. (2019). In "Civil unrest detection" there were two works: one by Ning et al. (2016) and other by Wei et al. (2020). There was no prevalence of any year in these two areas with a greater amount of work.

The last four areas had only one work each, in specific years: Al-Nabki et al. (2020) in the communal area "Drug-related crimes detection and Weapons' trafficking detection"; Saini and Bansal (2019) in Weapons' trafficking detection; Correa et al. (2018) in "Armed conflicts solution"; and Gil et al. (2018) in "Violence against woman analysis". The following section will present the details of techniques related to text mining found in the selected literature, crossing these techniques with the areas to determine the most recurrent ones by application area.

3.2 Most frequent techniques and technologies by application area

This section will separate the most frequent terms referring to techniques and technologies according to each application area in public security.

For "Cybersecurity", the most frequent technologies were Python and R programming languages, the Scikit-Learn and Natural Language Toolkit libraries, and the WEKA software. Regarding text mining techniques, those associated with machine learning are highlighted: Support Vector Machines, Decision Trees, Random Forests, and Naïve Bayes as the four most recurrent. There is also an emphasis on "term frequency-inverse document frequency" as one of the most frequent terms in the overall count.

In the "General crime detection/prediction" area, the most frequent technology was the Python language followed by the Natural Language Toolkit and RapidMiner, the latter a software dedicated to data mining. For techniques, those referring to machine learning were led by Naïve Bayes, followed by Support Vector Machines, Decision Trees, and Latent Dirichlet Allocation, to mention the most recurrent of this type. Term frequency-inverse document frequency shares with Naïve Bayes the rank of the second most frequent term. All other terms after the seventh position had only one occurrence for this area.

In the "Fraud detection" area, the most frequent technologies were Python, MatLab, and Scikit-Learn, and this time, the positions occupied were lower: Python sharing the sixth position with Random Forests and Named Entity Recognition; and MatLab and Scikit-Learn sharing the seventh position with eight other terms. Among the technologies mentioned, only MatLab is new compared to what has already been analyzed, being a proprietary platform for numerical analysis involving a dedicated programming language with the same name. Four machine learning techniques stand out: Support Vector Machines, Logistic Regression, Naïve Bayes, and Decision Trees, occupying the first, second, third, and fourth positions. Also, all other terms after the seventh position had only one occurrence for this area.

The terms extracted for "Terrorism detection" demonstrated a different trend compared to the areas previously explored concerning the frequency of technologies: the R language stands out, sharing the first position in the list with the techniques Naïve Bayes and term frequency-inverse document frequency. The other technology that appeared among the most frequent terms was Scikit-Learn, a library of techniques referring to the Python language. As mentioned, among machine learning techniques, Naïve Bayes appears first,

followed by Support Vector Machines, Random Forests, and Decision Trees. Term frequency-inverse document frequency, while a natural processing language technique remains in the spotlight.

For the "Cyberbullying detection" area, two machine learning techniques are the first: Naïve Bayes and Support Vector Machines. In the second position, along with the Python language and the term document-frequency inverse frequency technique, another machine learning technique can be found: the *k*-Nearest Neighbors. Three other technologies appear, sharing the third position: LIBSVM, an open-source cross-platform library that can be used with several other data analysis tools; Scikit-Learn and RapidMiner.

Tables A6, A7, A8, A9, A10 and A11 in Appendix A contain some counts about techniques and technologies in the previously commented areas. The following paragraphs continue presenting details about the remaining areas, but there are no related tables since the number of records found is much lower than in the previous areas.

Latent Dirichlet allocation appears as the most frequent machine learning technique in the "Digital / Cyber forensics" area, followed by Support Vector Machines, which shares second place with Natural Language Toolkit, Python, and Named Entity Recognition.

The "Support to the Judiciary power" area presented only term frequency-inverse document frequency as a recurring term, appearing twice. All other terms were unique, with emphasis on technologies: Python, R, Perl, Gensim library for natural language processing (for use with Python language) and WinBugs software, dedicated to Bayesian statistical analysis applying the Markov Chain Monte Carlo technique, which also appears among the extracted terms for this specific area. Naïve Bayes and Named Entity Recognition, which have been frequent in all other areas explored so far, also deserve to be mentioned for this area.

The "Support to Law Enforcement agencies actions" area has four terms referring to the most frequent techniques and technologies, each appearing twice: Latent Dirichlet Allocation, R language, Document-Term Matrix, and Named Entity Recognition. The term Document-Term Matrix refers to a matrix containing the frequency of terms in a document, being a relevant textual data structure, for example, for the application of topic modeling techniques.

The "Crimes' victims support" area has a set of more frequent terms that are somewhat different from the previous areas: General Architecture for Text Engineering, Java Annotations Pattern Engine, Rule-Based Language Expression Patterns, dictionaries, and manual annotation, each one appearing twice. General Architecture for Text Engineering, Java Annotations Pattern Engine, and Rule-Based Language Expression Patterns were used in Karystianis et al. (2018) and Karystianis et al. (2019), the first being a text annotation and categorization framework; the second, a pattern comparison language used in combination with the first; the third, an implementation made for application in the knowledge-based system developed by the authors, based on term dictionaries.

The "Sex-related crimes detection" area did not have any term more frequent than the others, all of which presented only one occurrence. Noteworthy is the presence of the R language, the open-source and cross-platform library LIBLINEAR, containing linear classifiers such as Support Vector Machines and Logistic Regression. Among the techniques: *k*-Nearest Neighbors, Logistic Regression, and Support Vector Machines.

The same occurs in "Drug-related crimes detection", not having terms with more than one occurrence. Here they highlight technologies such as Python and R languages, Automap, and ORA software, the latter two used by Neumann and Sartor (2016). While Automap

software is dedicated to text mining tasks, ORA is used for network construction. The presence of the latter software is consistent with the term Semantic Network Analysis found in the set referring to this area. Naïve Bayes, Logistic Regression, Support Vector Machines, and Random Forests also occur in this area.

In "Espionage detection", two techniques had two occurrences each: manual annotation and dictionaries. Fisher's Exact Test once appeared, referring to a well-known statistical test to compare different proportions for categorical variables. In the "Information security" area, no term was often higher than one. Here, the Python language and the Natural Language Toolkit can be mentioned among the technologies once again. A new technology emerged: the Openhangul API, used by Lee et al. (2016) for applying sentiment analysis.

In the "Software piracy detection" area, the only technology mentioned was WEKA, appearing only once. Of the machine learning techniques that recurrently appear in previously explored areas, only Random Forests occurred, and only once. Two other techniques that occurred once were: Weighted-Clustering-Coefficient and Graph-Based Static Birthmarking, both used by Sarwar et al. (2019). The Weighted-Clustering-Coefficient was applied to compute the similarity percentage between different computer programs under analysis, supporting the creation of the nodes in the Graph-Based Static Birthmarking technique.

Several techniques and technologies already presented were repeated in the “Civil unrest detection" area, but each with only one occurrence. Among the technologies, the Python language and its Gensim library stand out. Again, machine learning techniques stand out among those found: Naïve Bayes, Logistic Regression, Support Vector Machines, Adaptive Boosting, Extreme Gradient Boosting. There is an occurrence for the Word2Vec technique that consists of creating a textual features model to train classification models. All these techniques were employed by Wei et al. (2020).

The combination of areas "Drug-related crimes detection, Weapons' trafficking detection", as mentioned previously, presented only one related work, from which all terms were extracted. Al-Nabki et al. (2020) used the Python language and the Keras neural network library (written in Python), applying the Local Distance Neighbor Algorithm to recognize named entities that may be related to drug trafficking and illegal weapons sale through the Darkweb.

There was only one technology in the "Weapons' trafficking detection" area: the R Language, with one occurrence. Techniques such as Boosting, Decision Trees, Random Forests, Support Vector Machines reappear, each also with only one occurrence. There are new terms such as Cohen’s kappa Coefficient and Maximum Entropy, both with one occurrence and used by Saini and Bansal (2019). Cohen’s Kappa Coefficient was applied to check the agreement between two experts, and Maximum Entropy was used along with the other techniques mentioned earlier in this paragraph to perform classifications.

For the last two areas, "Armed conflicts solution" and "Violence against woman analysis", few techniques were detected, all with only one occurrence, since there was only one study related to each of them. In both areas, the only technology that occurred was the R language. In the "Armed conflicts solution" area, two terms related to techniques occurred: Document-Term Matrix and SMOG Index Calculation. The SMOG is an index related to the fact that adults use words in writing in a richer and more complex way than children. There are two other terms in the "Violence against woman analysis" area: Sentiment Analysis and Term Frequency. Figure 2 is a word cloud showing the most frequent word associated with

the term related to technologies and techniques from the selected literature, separated by unigrams.

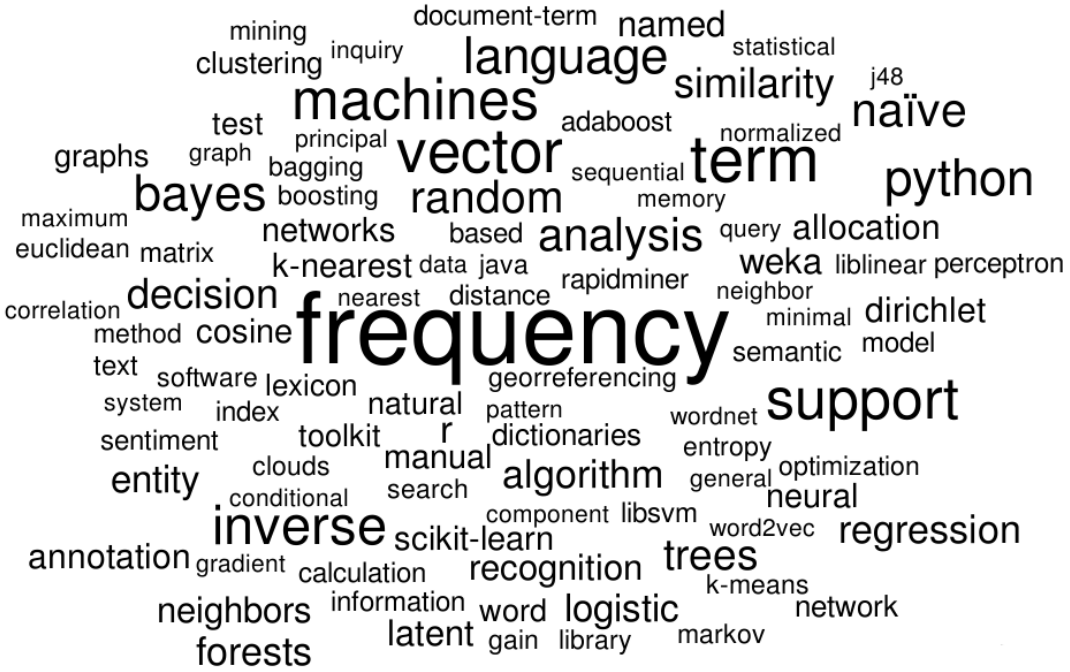

Figure 2. Wordcloud with the most frequent words in the technologies and techniques related terms.

In Figure 2, the central and most frequent word is “frequency” both associated, for instance, to “term frequency/” and “term frequency-inverse document frequency”, very recurrent statistical measures for words counting in text mining. In a second level, term as “support”, “vector” and “machines”, “naïve”, “bayes”, related baseline machine learning techniques appear, as well as “python”, the most recurrent programming language in the selected literature. The following section will address the opportunities and challenges for text mining in public security found in the selected literature.

3.3 Detailing the opportunities and challenges for text mining in public security

For each of the areas where there were indications of future work, in the list of 92 articles that contain this type of indication, extracted from the more extensive set of selected literature, a synthesis of these indications will be presented to identify opportunities and challenges for the outline a research agenda.

It should be noted that not all 19 identified areas have works with mention of future directions to be presented in the construction of this part of the text; thus, following, the seventeen areas with works containing the desired information are presented.

3.3.1 Cybersecurity

The cybersecurity application contained the largest number of further development and research proposals, with twelve directions indicated in 2018, eight in 2019, and eight in 2020.

Elkhawas and Abdelbaki (2018) presented an approach using trigrams and portable executable file attributes as features for malware detection, suggesting the automation of the mining process and using other techniques than those previously used. They also commented on using support vector machines with larger datasets to improve accuracy by including more features, performing more experiments, and trying to group malware in families.

Hadad et al. (2018) also developed research dedicated to malware detection, suggesting using more features in the malware detection process through user-generated content instead of libraries or sets of codes. Ariffin et al. (2018) built a ransomware background knowledge base and indicated the need to extract more features using texts to identify ransomware families.

Silomon and Roeling (2018) mined people's opinions about the concept of "Software as a Weapon" by linking traditional aspects of weapons to understand the differences between software and malware used from an international security perspective. They suggested reducing the questionnaire to collect the opinions by selecting the most informative questions according to the highest factor loading for the related components.

Chung et al. (2018) developed an approach to identify social media resources correlated to abnormal stock returns from vulnerable companies targeted by cognitive hackers. They indicated that the experimental results of the research would be open for researchers interested in applying their approach to obtain additional evidence for its effectiveness.

Dong et al. (2018) proposed a framework for discovering new cyber threats in darknet marketplaces with new research ideas on how to extract information from short texts (also considering foreign languages) in the dark web, obtaining more features for classification, and applying natural language processing to increase the performance of the framework.

Concepción-Sánchez et al. (2018) proposed a text mining and fuzzy logic system to detect compromised user accounts. They suggested incorporating other methodologies to improve the system's decision-making and indicated the development of a mobile application using a theft detection system like the one they proposed.

Husari et al. (2018) developed an approach to automate the extraction of low-level cyber threat actions from publicly available intelligence sources to enable timely defense decision-making. Their indications for future works were related to the extension of the approach for viewing threat action as a verb-object representation to other syntactic blocks and proposed a new way to automatically parse all types of threat action expressions by extracting them to provide vital information for more analysis.

Ruano-Ordás et al. (2018) developed an empirical study about concept drift to discover its origin, types, and undesired effects in the context of e-mail classification. Their future research ideas were related to measuring the impact of the proposed methodology in various scenarios, developing new methodologies to estimate the performance of classification systems without changing the underlying concept drift, and designing new classification systems to detect specific types of concept drifts in real-time.

Zainal et al. (2018) produced a tool to aid users in distinguishing potentially severe spams by measuring the risk in the content. They identified further directions by developing

an automated environment for new experiments with a prototyping dendritic cell algorithm and its deterministic version. They also suggested using a larger-sized corpus, checking its consistency and reliability, and applying new simulations to validate the severity concentration assessment in other spam forms.

Sonowal and Kuppusamy (2018) presented an anti-phishing model, entitled Smishing Detection based on Correlation Algorithm, and indicated the use of a new feature selection algorithm to obtain more varied features. They also suggested exploring deep learning techniques and enhancing the prototype model's user interface as future directions.

Samtani et al. (2018) presented an approach to identify Supervisory Control and Data Acquisition (SCADA) devices through a search engine to find computers connected to the Internet and then assess the vulnerabilities of the devices with a state-of-the-art tool. They suggested the new work may be interesting for identifying specific device owners and locations, categorizing the devices in a fine granulation to understand vulnerabilities that affect some subsets, and employing assessments at regular intervals to know how SCADA vulnerabilities evolve.

Balim and Gunal (2019) proposed a machine learning-based model to detect smishing messages in short message services. They indicated that for future studies, one way forward is to apply smishing message analysis in other languages, using different features and classification algorithms.

In their study, de Boer et al. (2019) proposed the Horizon Scanner tool, based on crawling and scraping relevant text databases on potential threats and trends in cybersecurity to speed up forecasting. The authors indicated that future interactions with their tool would emphasize human-centered aspects, such as workflow support and greater explanation and controllability of the analyses.

Gravanis et al. (2019) proposed a model for fake news detection using content-based features and Machine Learning (ML) algorithm. As advances in their original proposal, they highlighted the use of multiple metadata about the source and author of the news, along with information dissemination capabilities on social media and the use of deep learning methods with larger datasets to improve detection of false news.

Kakavand et al. (2019) proposed the anomaly detection system called Online Adaptive Deep-Packet Inspector to classify web service message attacks. As a perspective for further developments on their proposal, they indicated the possibility of analyzing larger volumes of information to ensure an effective scaling of the detections for which the system was developed.

Mohasseb et al. (2019) investigated how the data from multiple companies representing different incidents can improve classification accuracy by aiding the classifiers to identify different types of incidents. The use of other datasets and machine learning techniques and the investigation of class imbalance to improve the classification accuracy were directions identified for future research.

Roopa and Induja (2019) proposed a framework based on text mining to avoid the risk of opening suspicious e-mails and suggested further work in the customization of this framework by improving the classification of sentiments towards the online content and incorporating library files for selective predictions.

Sudha and Rupa (2019) developed a framework to classify and match the various cybercrime incidents using the Naïve Bayes classifier. They comment that advances in the system can be implemented, ensuring that preventive measures are also suggested based on the results.

Palad et al. (2019) seek to determine whether predictive text mining analytics can be used effectively to identify and classify online scam data to gain insights and help uncover patterns in a cybercrime dataset that assists in criminal investigations in the Philippines. They highlight that developing or obtaining more inclusive or relatively large cybercrime datasets in future studies is necessary to significantly improve relevance of the results.

Alagheband et al. (2020) analyzed cybersecurity trends to propose a conceptual framework for identifying cybersecurity topics of social interest and emerging topics that need to be addressed by researchers in the field. Three considerations can be pointed out about future studies based on the authors' comments: (i) apply gap analysis of cybersecurity trends between newspaper corpora and cybersecurity whistle-blowers to reveal some noteworthy points; (ii) explore the correlation between cybersecurity-based standards and patents with either academic papers or the whistle-blowers' reports to yield new insights to be studied, for instance, using causality analysis; and (iii) use the proposed framework for other applications or even some specific and targeted sections of cybersecurity.

Al-Ramahi et al. (2020) extracted topics of interest from hackers' websites to use them as security controls and systems inputs. They suggest integrating these topics of interest as a new category of commitment indicators for application security controls and systems for future studies.

Battaglia et al. (2020) proposed a way to evaluate the harmfulness of any form of content by defining a new data mining task called "content sensitivity analysis". They define two directions for further studies: (i) investigate more complex techniques borrowed from machine learning, computational linguistics, and semantic analysis; and (ii) build massive and reliable annotated corpora to ensure that the performance of any content sensitivity analysis tool is sufficient, no matter how complex the learning model.

Calderon et al. (2020) compared the performance of decision tree classification algorithms using a large dataset to assess whether such algorithms can be used effectively in data mining efforts to positively impact or contribute to criminal investigations. They highlight two possibilities for future studies: (i) compare the results obtained from the classifiers with other decision tree classifiers implemented in WEKA and investigate what causes the difference in the performance of such algorithms; and (ii) perform weight allocation, subsequently applying rank approaches to ranking classifiers.

Ma et al. (2020) presented a data-driven approach, based on text-mining methodologies, for classifying transient events and identifying fake events caused by false data attacks in power systems. These authors separated three directions for the development of new studies: (i) include a more diversified set of power system events for classification purposes, such as the switching of synchronous motors; (ii) investigate other types of data spoofing strategies to provide more information on applications of the proposed approach; (iii) use other useful information recorded by phasor measurement units to improve the accuracy and robustness of the proposed approach.

Palad et al. (2020) significantly improved their previous study (Palad et al., 2019) based on the identified research problems: relative small datasets and different classification methods. They suggested comparing the results obtained from these decision tree algorithms by applying appropriate weight allocation and subsequent ranking approaches to evaluate those classifiers. They also commented that the classification of cybercrime data could be conducted using other available data mining tools or techniques.

Samtani et al. (2020) developed a cyber threat intelligent detection framework called "Diachronic Graph Embedding Framework" to explore online hacker forums to identify

emerging threats in popularity and tool functionality. The authors determine that the framework can be employed in other related areas, such as improved memory forensics and malware evolution on datasets extracted from VirusTotal (an online URL analysis service to identify malicious content).

Pires and Georgieva (2020) developed an intelligent tool for phishing detection based on machine learning that integrates only local information and does not rely on blacklists or other external sources. The authors commented that implementing the proposed tool in commercial environments would be favored if embedded into a platform for online message labeling. They also pointed out that periodic retraining with recently accumulated labeled data over a limited time would guarantee a smooth adaptation to new attack trends.

3.3.2 General crime detection/prediction

The "General crime detection/prediction" term was the second set of applications that included further development indications, with six in 2018, five in 2019, and five in 2020. The research by Aghababaei and Makrehchi (2018) (i) verified the contribution of content extracted from Twitter in a crime prediction model, (ii) identified an efficient time forecast model that captures the most relevant resources, such as time topics for forecasting, and (iii) explored the temporal effects of the content of crime directions to determine the lag between the predictive features and the crime trends. They proposed for future research working on a sampling approach to avoid the loss of user activities and collect data from historically active users to allow the use of text semantic analysis to better understand the relationship between resources. Additional analysis incorporating other socioeconomic and geographical information correlated with information on criminal activities was also suggested.

AL-Saif and Al-Dossari (2018) intended to detect crimes, identify their nature using different classification techniques, and evaluate their speed and accuracy performance. They suggested including a spatial and temporal analysis to determine when and where crime has spread earlier to help predict how it is likely to spread in the future.

Percy et al. (2018) identified regions for which data had internal similarities and could be combined for creating a reliable crime prediction model. They suggested incorporating a time-series into the developed model by using, for instance, Long Short-Term Memory modeling of sequential data.

Po and Rollo (2018) analyzed and mapped thefts through online newspapers using text mining techniques for an Italian city. The following were recommended for future developments: (i) seek collaboration with law enforcement agencies to deal with real-time crime data, (ii) perform analysis and integration of crime news from other local newspapers, (iii) remove news duplicates, (iv) represent more types of crimes, (v) apply topic detection, (vi) apply advanced crimes analysis, and (vii) transform the stored information in linked open data.

Ristea et al. (2018) explored the spatial relationship between crime events and nearby Twitter activity around a football stadium by estimating the possible influence of tweets for explaining the crimes' occurrences in the neighborhood on match days. The suggested directions for further research were (i) perform seasonal analysis of crime patterns by increasing the database for performing comparisons, (ii) expand the study space to include areas that may also experience changes in crime according to different conditions caused by large-scale events, and (iii) model variability in crime volumes on weekdays compared to weekends as dependent variables.

Tran and Tran (2018) proposed techniques for aspect extraction and sentiment analysis in aspect-based opinion mining problems with applications in criminal activities analysis. They indicated further research to improve proposed techniques performances, solving the problem of determining the strength of opinions, analyzing opinions expressed with adverbs, verbs, and nouns, and developing more research and experiments in political and security domains.

Mine et al. (2019) investigated if crime messages sent by e-mail can be further explored as a valid source for analyzing the criminal characteristics of a region. They proposed employing other machine learning techniques, particularly Gradient Boost Decision Trees and Deep Neural Networks, to develop a method for automatically extracting essential features from the crime reports. They also suggested developing methods for analyzing and tracking temporal and spatial changes in crime and predicting local crime.

Das and Das (2019) proposed a graph-based clustering technique for discovering labels of crime reports based on extracted paraphrases from large untagged crime corpora. They determined the need for developing an automatic system for labeling the crime reports based on extracted paraphrases.

Malim et al. (2019) studied annotated sentiments in a list of Malaysian Tweets possibly subject to crimes or illicit messages from the stance of the psychotic trait to recognize criminal activities using machine learning. They highlighted as a possible evolution of their study “the use of more data to predict criminality texts by incorporating artificial intelligence and deep learning concepts”.

Nedeljkovic et al. (2019) developed a study about the similarity measure that provides the most precise results about the crime domain in a question-answering system. They specified the need to develop a new system that will not be necessary to involve an expert.

Savaş and Topaloğlu (2019) examined the wisdom of crowds on crime issues from social media to illustrate the relationship between social whispers and the current crime situation in Turkey. Changes in keywords, filters, and target groups applied to the same analysis in other fields were indicated for future research.

The study by Alatrista-Salas et al. (2020) intended to determine what types of crimes were perpetrated based on the news. They defined four points for further studies: (i) extend the applied methodology to an online platform that can process and analyze streaming documents; (ii) answer the questions about who commits a crime and where to enrich the methodology applied to build a comprehensive approach; (iii) extend the methodology applied to other languages and criminal texts; and (iv) improve points such as time, memory efficiency and the ability to build a classifier capable of detecting more than one crime in a given news article.

Birks et al. (2020) sought to identify specific crime groups from unstructured free texts containing *modus operandi* data within a single administrative crime classification. The authors defined four directions for future research based on their study: (i) to assess the applicability of the latent Dirichlet allocation for free text associated with other crimes; (ii) the use of other topic modeling techniques besides the latent Dirichlet allocation, allowing the incorporation of additional variables in the model; (iii) compare and contrast the space-time structure of different types of assault modus operandi as identified through the model; and (iv) use the probability of allocation of topics to documents to identify new or emerging modus operandi of crimes from documents that are not well categorized by latent Dirichlet allocation when trained against an existing model, supporting the development of early warning systems that can identify emerging criminal behavior.

Mukherjee and Sarkar (2020) proposed a system that automatically extracts crime locations from huge newspaper collections to get the picture of high crime-prone locations in a given locality. They define two directions for future studies based on the proposed system: (i) use more training data for tagging crime information with Conditional Random Fields, and (ii) apply deep learning to improve each system module.

Lal et al. (2020) analyzed Twitter data to identify tweets about crimes that need police attention, proposing an approach to doing so. They define five directions for further studies on the topic they worked on: (i) using more classifiers to test the effectiveness of the proposed approach; (ii) adding location information to the crime tweet to help police identify the crime location; (iii) apply a set learning-based approach to crime tweet classification; (iv) apply natural language processing techniques such as grammar class markup to improve resource extraction; and (v) carry out a comparative analysis of the approach they used with the other approaches.

Saldana et al. (2020) presented a methodology for analyzing criminal facts in online newspapers, identifying the different communes where the greatest number of criminal events occurred. They suggested the use of a more significant number of sources such as social networks, the breakdown of events by limited geographic areas (streets or sectors of a given commune), and the incorporation of techniques such as sentiment analysis to study in more detail at levels of violence associated with specific events.

3.3.3 Fraud detection

Works on "Fraud detection" applications with comments about further research appeared in two papers in 2018, three in 2019, three in 2020, and one in 2021. Bhardwaj and Gupta (2018) proposed a text mining framework for detecting financial statement fraud by identifying and analyzing the linguistic data available in financial reports. They suggested future work implementing the proposed framework using a dataset of fraudulent and non-fraudulent financial statements and text mining tools.

Lee et al. (2018) presented a new forward and backward analysis methodology implemented in an Information Extraction prototype system. The following future research directions were derived based on the limitations of the developed work indicated by the authors: (i) work on a keyword extraction template for other financial crimes on financial discussion boards, (ii) use other data artifacts, such as "broker ratings" and "director deals," for the forward and backward analysis, and (iii) program the prototype system to obtain comments through HTML files by crawling comment data from financial discussion boards.

Dastjerdi et al. (2019) applied text mining to detect high fraud risks in companies to compare their precision. Their set of techniques included: F-Limer and Hausman tests to determine the type of pooled data; the modified Wald test for heteroscedasticity in the fixed-effect regression model; the Wooldridge test for autocorrelation of residuals in the panel data; Convex Optimization, Minimum Absolute Reduction, and selection operator regression techniques to perform fraud risk detection. Their future research indications included (i) the use of other methods, such as neural networks, genetic algorithms, decision trees, Bayes analysis, and fuzzy analysis, for comparison purposes, (ii) the use of other text, such as an auditor report, (iii) increase the periods for fraud detection on the analysis, and (iv) develop a new analysis for different languages and in other countries.

Rabuzin and Modrušan (2019) compared prediction models using text-mining and machine-learning techniques to detect suspicious tenders in public procurements and to apply

these techniques to develop an approach to detect suspicious one-bid tenders. They comment that it is necessary to include additional indicators to increase model accuracy and apply neural networks, deep learning, and other machine learning algorithms for better results.

Sahu et al. (2019) proposed an architecture for credit card fraudulent transaction detection using machine learning. The authors defined that new studies can be developed on the attribute name not provided in the applied dataset. They also suggest that different classifications can be performed using Greedy Search Optimization and genetic algorithms to get better results. Collecting a more extensive dataset to improve predictions is another direction the authors set for further studies.

Angenent et al. (2020) applied machine learning to perform explainable business sector predictions from financial statements with fraud detection applications. The authors highlighted as a perspective for future work “using other classification algorithms, optimized hyperparameters, and different data sources, such as written annual reports (for text mining)”.

Kabwe and Phiri (2020) presented a metrics model based on distance metrics to quantify the credential identity attributes used in online services and activities to detect identity theft. They define as a continuation of their research the implementation of the results obtained for constructing a multimodal solution, consolidating previous work in this area, arriving at a single and robust solution capable of recognizing how much threat must be eliminated in services and online activities.

Monish and Pandey (2020) applied a comparative analysis of data mining models for fraud detection to predict fraudulent firms. They provided three directions for future work: (i) compare advanced deep learning algorithms for text classification, such as recurrent neural networks; (ii) explore new advanced ensemble techniques to make further comparisons of models; and (iii) explore data mining techniques other than text, for example, analyzing multimedia data such as audio, video or images.

Siering et al. (2021) developed an artifact that can act as a robust classifier by providing an assessment of whether a given document is suspected of being fraudulent applying it to the stock market. The authors indicated several directions that could be taken in future studies and developments: (i) the inclusion of the proposed fraud detection classifiers in a fraud detection system to improve the "information-based market manipulation detection capabilities" of companies and market surveillance authorities; (ii) the use of classifiers to complement established fraud detection systems covering other manipulation scenarios; (iii) the inclusion of classifiers in browser toolbars, which already generate warnings for phishing sites; (iv) the use of design principles and design features to improve the robustness of the classifier by applying it to other fields or languages; and (v) to investigate the robustness of text-based classifiers in the social commerce context.

3.3.4 Terrorism detection

Terrorism detection applications presented eight works with future research recommendations, with five in 2018, one in 2019, and two in 2020. Al-Khalisy and Jehlol (2018) worked on a method for detecting tweets with terrorist ideas and identifying the location of people who own these posts and proposed applying their method to other fields.

Alguliyev et al. (2018) proposed a method for detecting terrorism-related activities in e-government environments and proposed further experiments based on the method compared to other approaches. Mansour (2018) examined what eastern and western people think about ISIS and investigated if there is a significant difference between the proportion

of negative and positive words within users' tweets from both regions. They suggested repeating the study to include more countries to collect more tweets for generalizing the results and collecting multilingual tweets for analysis in different languages.

Öztürk and Ayvaz (2018) investigated the public opinions and sentiments towards the Syrian refugee crisis and suggested including more languages for further study. Zahra et al. (2018) presented an approach applying machine learning algorithms to identify extremism-related content and users. They proposed, for future studies, the use of demographic information, user activity feeds, and links between users to detect the location of extremists.

Petrovskiy and Chikunov (2019) proposed an approach for detecting radical users in social networks by analyzing their relationships and features as vertices on a social graph without using textual content they generate. Their recommendations for further analysis included (i) applying an extension of the label propagation approach by imitating topics spreading between users based on their relationships, (ii) including this information for radical members of social network detection to improve quality, and (iii) predicting the probability of occurrence of communication between users in time to prevent extremist activities.

Castillo-Zúñiga et al. (2020) proposed a methodology to obtain added value from datasets extracted from hundreds of downloaded web pages and developed a software architecture to test this methodology, applying it to cyberterrorism vocabulary detection. The authors pointed out the need to carry out studies related to messages that include terrorist content on Facebook and Twitter, using sentiment analysis and opinion mining techniques to extract people's comments and classify the emotional tone.

Miranda et al. (2020) studied radicalism intention using content detection. They pointed out a direction for future studies based on sentiment analysis techniques to improve content detection and classification performances.

3.3.5 Digital/Cyber forensics

Two works included future research recommendations in Digital/Cyber forensics applications: one from 2018 and the other from 2019. Giacalone et al. (2018) opened a debate on the study and use of statistical and computational methods for web data on new forensic topics, proposing a system for predictable jurisprudence using automatic sentence analysis. The authors also proposed, for future research, dealing within a Big Data context by considering the map/reduce approach with a real estimation of the evolution based on the solution they presented to analyze administrative judgments.

Henseler and Hyde (2019) proposed using a graph database and query language to assist in answering key digital forensic investigation questions. The authors defined three new study lines based on their work: (i) study how forensic investigators can interact with the graph generated by their approach and how to extend the graph with other data sources; (ii) study how events on a timeline can be added to the graph, how non-digital evidence information can be included, and how to improve the performance of existing entity extraction techniques in unstructured email and documents; and (iii) study whether new machine learning techniques, such as graph neural networks, can be used to learn from investigators which link and event patterns are interesting from the investigator's point of view.

3.3.6 Cyberbullying detection

Nine works with future development indications were found in the cyberbullying detection application area: two in 2018, one in 2019, four in 2020, and two in 2021. Alakrot et al. (2018) built an offensive language-based database to detect antisocial online behavior. They presented an exploration of more pre-processing techniques and machine learning algorithms to create better offensive content detection models in Arabic online communication as continuity for their research.

Ventirozos et al. (2018) presented a new approach using the sentiment analysis at the message level but considering the entire communication thread as the context of the aggressive behavior. They proposed investigating the representation of sentiments as word embeddings learned through deep neural networks as a future development for their research.

Andleeb et al. (2019) proposed a text mining framework to detect bullying text proactively. They commented about using specific adaptations to make the framework data-independent and to use it for any other dataset for future works. They also commented about adding more bullying words in the bullying words list to make the study suitable for real-world problems. As the last direction, they defined applying sarcasm detection to further improve the framework's performance.

Adikara et al. (2020) studied cyberbullying detection on Instagram comments divided into two classes: cyberbullying and non-cyberbullying comments. For future works, the authors defined the exploration of methods such as support vector machines or neural networks to verify their detection rates, demonstrating that these methods are suitable for the application in cyberbullying detection.

Ishara Amali and Jayalal (2020) introduced an automatic approach to detect Sinhala language social media comments to insult a person. For further research, they pointed out: (i) expanding the data corpus to include Sinhala language comments that are also written using English characters; (ii) improving the applied model to obtain high accuracy with increasing recall value with a larger data corpus; (iii) the expansion of the Sinhala swearword list, with the help of the participants in the labeling system and with the knowledge of experts; and (iv) the automation of the manual labeling process to make the process more efficient.

Slamet et al. (2020) analyzed text documents on social networks and then classified them into two classes: one with indications of bullying and the other clean concerning these indications. They proposed two advances over what they had already developed: (i) add more documents to increase accuracy; and (ii) use the supervision of a linguist and manually check the dataset that will be used in the training and testing process so that the documents to be processed are entirely free from noise and other errors.

Wang et al. (2020) proposed a multimodal detection framework using multimodal information in social networks to deal with cyberbullying. For future developments, they defined the use of multimodal information fusion in cyberbullying detection, considering modal data associations in social networks, supporting the modeling of new types of cyberbullying behavior.

Bozyiğit et al. (2021) tested the classification of online bullying content by using social media variables with classical text mining approaches. As proposals for future work, they highlight: (i) create datasets belonging to different countries/regions that use the same language to compare the effects of social media demographically; and (ii) implement a fuzzy rules-based system for current approaches using predictions based on social media.

Choi et al. (2021) proposed a practical method of identifying social network users who make high rates of insulting comments. For research purposes, the authors defined their method as a basis for further studies since no methodologies are dedicated to identifying key cyberbullies. The authors consider using the proposed method to classify benevolent comments for practical purposes. They also pointed out that identifying and ranking key cyberbullies can help online platform operators mitigate cyberbullying.

3.3.7 Support to law enforcement agencies actions

Among the works related to supporting law enforcement agencies, four presented directions for further research: both 2019 and 2020 had two works. Basilio et al. (2019) developed a methodology for knowledge discovery in emergency response service databases based on police occurrence reports to support law enforcement agencies planning to investigate and combat criminal activities. They appointed interactions with multi-criteria methods to support decision-makers choosing actions to identify illegal demands as activities for future research.

Behmer et al. (2019) designed a systematic approach for the processing and extracting formalized semantic concepts to assist criminal investigations and suggested using instantiated information, such as temporal sequences, to support semantic reasoning about the occurrences of events and assist in criminal identification.

Basilio et al. (2020) developed a method for knowledge discovery in emergency response databases based on police incident reports. From their research, they were able to make the following indications about future studies and developments: (i) apply the optimization of material and human resources in the application of policing strategies; (ii) improve the modeling using other multicriteria decision aid methods, such as PROMETHEE V; and (iii) and carry out research expansions in other countries to consolidate results on the impact matrix of policing strategies.

Hou et al. (2020) proposed a “Bidirectional Encoder Representation from Transformers” based on the Chinese language relation extraction algorithm for public security, which can effectively mine security information. They proposed further exploring the application of deep learning in relational extraction and developing the model for public security and the Chinese language structural analysis.

3.3.8 Sex-related crime detection

For the sex-related crime detection applications, only one work suggested further research. Hultgren et al. (2018) proposed an information system approach to identify sex trafficking victims based on analyzing online ads and specified the following directions to improve the system: (i) examine male victims to see if the indicators and keywords are different compared to female victims, (ii) investigate technologies to create an automated knowledge management system on the topic, (iii) develop an automated data extraction process, (iv) use data from online services for adults to develop studies considering the sex industry, and (v) determine the meaning of emojis in texts to assess if they can be used to detect victims of sex trafficking.

3.3.9 Support to the Judiciary power

The support for judiciary applications counted three works with future research directions. Iftikhar et al. (2019) developed a legal mining system based on machine learning algorithms to recognize named entities on judgment texts. They indicated for directions of further work (i) the inclusion of other types of judgments to be tagged and used in training and testing of algorithms, (ii) the preparation of a comprehensive dataset, containing a variety of court judgments, (iii) the application of various deep learning algorithms for named entity recognition and preparation of specialized pre-trained word embedding for legal text on legal text, and (iv) providing a variety of applications on extracted named entities.

Pina-Sánchez et al. (2019a) reported the results of an analysis on a new sentencing database to explore if offenders with common Muslim names attract a different sentence and if any notable differences can be attributed to discriminatory sentencing practices. The authors highlighted the following elements to be considered in new research: (i) convert sentences to a reduced and censored form for use by specialists in other areas who are interested in performing analyses on these sentences, (ii) link the Sentencing Council data to the Ministry data by capturing the ethnic background of the defendant, (iii) conduct multivariate analyses to determine if there is evidence of ethnic discrimination in sentencing, and (iv) include all decisions taken before the final sentence for analysis.

Xia et al. (2019) evaluated the effect of judge gender on judicial decision-making in China. They indicated the need for further study on the interaction between gender roles, legal and extra-legal factors in judicial processes.

3.3.10 Drug-related crime detection

The work by Cichosz (2018) was the one identified with new research directions in drug-related crime detection applications. The authors developed a study applying classification algorithms to categorize forum posts as drug-related. As advancements to their work, they suggested (i) to include more selection algorithms for text mining purposes by comparing their performances, (ii) to explore enhancements for the process of deriving bags of words and the text representation referred on the article, (iii) to incorporate additional non-text attributes in the representation of forum posts by trying to improve the quality of the classification, and (iv) use manually annotated class labels to represent the discussion topics.

3.3.11 Espionage detection

Only one work with directions for future research occurred for Espionage Detection, published in 2020. Glancy et al. (2020) analyzed the association between malicious insiders' characteristics and the different types of attacks on organizations. The authors commented that they would attempt to enlarge the sample size by adding new cases reported by the news media in future research. They also intend to focus more on the mechanism and overarching theories to explain the behavioral patterns of malicious insiders.

3.3.12 Crimes' victim support

Crime victim support applications presented indications for new research development in one paper from 2018, one from 2019, and two from 2020. Karystianis et al. (2018) examined if automatic text mining of domestic violence police event narratives is feasible for identifying mentions of mental health disorders at the narratives of people

involved by employing a knowledge-driven approach. They recommended for future research: (i) examine the authenticity of informal mentions of mental health disorders through formal diagnoses in administrative data collections, (ii) expand the set of information extracted from police narratives to assess the characteristics of people who committed violence and victims for risk groups, (iii) create predictive models to investigate if recurring domestic violence events are predictable for groups at risk by informing preventive strategies.

In a second study, Karystianis et al. (2019) took a similar approach as the previous article by investigating if the application of text mining can automatically extract abuse types and sustained victim injuries from a corpus of domestic violence events. Their recommendations for future research included: (i) use information from a corpus of domestic violence in combination with the collection of administrative data on mental illness to further examine the links between mental illness and domestic violence, exploring the relationship of types of abuse with gender and victim injuries, (ii) perform a new analysis by combining demographic variables with the results previously achieved, (iii) combine victim injuries extracted from clinical data resulting from contacts with health services to assist in identifying victim abuse and implementing intervention strategies, (iv) apply modeling to investigate if people with characteristics of interest can predict the severity of the abuse by assessing if specific victim phenotypes are prone to specific abuse.

In a third study, Karystianis et al. (2020) presented the prevalence of extracted mental illness mentions for persons of interest and victims in police-recorded domestic violence events, following a line of work that continues the previous two. The authors commented that information drawn from a police domestic violence dataset suggests that there may be more detailed information on mental illness trends in victims and persons of interest. This approach can provide the basis for examining the agreement of extracted mentions of mental illnesses with the official diagnosis of health records and surveys that aim to assess victims and person of interest characteristics in police records. There is also the idea that the information extracted can be used to design predictive models about the risk of further victimization, informing prevention strategies that can be implemented in the early stages of police involvement in a domestic violence event.

Lyu et al. (2020) applied sentiment analysis over textual data from Weibo social media to explore people's attitudes towards child abuse incidents, the reasons behind them, and how people's emotions will become the potential driving force for improving child protection policies in China. The authors consider applying machine learning in sentiment analysis as an ideal way to improve the accuracy of this type of analysis in textual data, indicating this possibility for future studies.

3.3.13 Software piracy detection

Software piracy detection applications appeared in two works: one from 2018 and another from 2019. Kim et al. (2018) proposed a software classification scheme for an efficient software filtering system, applying it to detect pirate software. They recommended devising adaptive techniques by applying static and dynamic characteristics to study the trade-off between analysis speed and detection of overshadowed programs. Sarwar et al. (2019) proposed a new software birthmarking approach based on hybrid text and graph mining techniques to support pirate software detection. Their recommendation for future work was to test a hybrid approach to software theft detection using graph-based birthmarks and software watermarks.

3.3.14 Drug-related crime detection and Weapons' trafficking detection

Here, there is a combination of two of the identified areas of text mining application in public safety, where there was only one work specifically in 2020. Al-Nabki et al. (2020) presented an end-to-end neural network architecture to recognize emerging and infrequent named entities in noisy user-generated text, supporting detecting suspicious activities associated with weapons and drug selling in Darknet. To continue their research, the authors considered representing the context of the input token by evaluating its Local Distance Neighbor vector using Bi-directional Long Short-Term Memory to see how this improves the method's performance.

They also pointed out using other space-embedding representations such as FastText or StarSpace to represent the training entities so that the cosine similarity can be measured between the input token and the training entities instead of the training tokens. As the last point, they commented on incorporating graphical resources extracted from images published in Tor domains, which would hopefully increase performance

3.3.15 Armed conflicts solution

Applications in armed conflict solutions contain one work by Correa et al. (2018), who proposed a methodological approach to show how to linguistically analyze peace agreements as acceptable political products based on the difficulty of the text. As a direction for future work, they commented on the need for analyzing peace agreements of other non-Spanish-speaking cultures and failed peace accords by adapting the script developed for the research they applied.

3.3.16 Weapons trafficking detection

Weapons' trafficking detection contained one work by Saini and Bansal (2019). The authors proposed a novel approach to identify the procurements of modern weapons over the dark web forums by terrorists. They suggested enhancing the illegal weapon procurement model to support different languages used by violent extremists in dark web forums and apply social network analysis. Another suggestion was to use topic modeling to discover more hidden patterns in the dark web and detect more illegal activities.

3.3.17 Civil unrest detection

Civil unrest detection is the final public security-related application area for text mining containing future work directions. Wei et al. (2020) examined the changes in the behavior of Twitter users regarding prejudice against immigrants following recent protests in the United States on topics related to immigration. The directions for further study were: (i) improve prediction of change at the user level with better-annotated data; and (ii) study a broader population in different countries and study the long-term effect of protests on online prejudice.

3.4 Comments on Ethics and the Use of Text Mining in Public Security

It is important to comment on the ethical question in using text mining in public security to close the discussion. In several cases of application of text mining techniques,

issues regarding ethics in using sensitive and personal data emerge, even though these data are necessary for some action aimed at people's security. Some articles made comments about the ethical use of data about people, victims of abuse, or crimes in the selected literature.

Hultgren et al. (2018) highlighted the problem of using data on victims of sexual exploitation. This was one of the main limitations of his research for accessing information since, for reasons related to research ethics, it was not possible to make direct contact with people in this condition. This fact alone demonstrates the delicacy of the matter. Neither the victims wish to expose themselves for their safety, nor do law enforcement authorities publish detailed textual records to avoid exposing these people, whether scientific research.

Karystianis et al. (2018), Karystianis et al. (2019), and Karystianis et al. (2020) accessed domestic violence case data to develop their studies. However, they commented about the ethical issue in using the data source (a police narratives base). A strict security protocol was applied, ensuring that the text mining of the cases' narratives could only be carried out *in loco*, at the headquarters of the police department in the region where the study was carried out, and only de-identified outputs could be extracted to be applied on their research. In summary, the use of text mining tools could be considered ethical if it is oriented towards maintaining social welfare, avoiding the exposure of people mentioned in texts, and following the information security regulations required by public security authorities.

4. Conclusion

The detail presented in this article brings all the selected literature, with the separations related to each of the items targeted by the research: (i) main application areas, (ii) techniques and technologies employed, and (iii) opportunities and challenges according to the most recent literature (from 2018 to 2021), looking at the directions or proposals for future research. While in (i) and (ii), the whole selected literature with 194 works was explored, in (iii), a subset containing 92 works was extracted. A publicly available repository was created in GitHub (see the footnote of Section 3.4 in the present text) and a spreadsheet with all information extracted from the works selected in the systematic review process.

Nineteen application areas related to public security were identified. The related literature was presented among Tables 1 and 10, with the last two tables grouping two or more application areas and subsequent text, presenting areas with smaller amounts of literature (one or two works only). After presenting the application areas, the most recurrent techniques and technologies applied were presented according to each identified area, with data being complemented with a series of tables in Appendix A.

Detailing about future research directions of the 92 most recent works is presented, separating them according to the seventeen application areas in which they were detected, describing these directions in a summary paragraph. A final comment was made regarding the ethical question in using text mining in public security, demonstrating the importance of ensuring that sensitive data, particularly about people, are kept confidential to avoid compromising their physical and social integrity, demonstrating some works that illustrate restrictions in this regard.

References

Aboluwarin, O., Andriotis, P., Takasu, A., & Tryfonas, T. (2016). Optimizing Short Message Text Sentiment Analysis for Mobile Device Forensics. In *Proceedings of the 12th IFIP International Conference on Digital Forensics* (pp. 69–87). https://doi.org/10.1007/978-3-319-46279-0_4

Adikara, P. P., Adinugroho, S., & Insani, S. (2020). Detection of cyber harassment (cyberbullying) on Instagram using Naïve Bayes classifier with bag of words and lexicon based features. *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*, 64–68. https://doi.org/10.1145/3427423.3427436

Aghababaei, S., & Makrehchi, M. (2018). Mining Twitter data for crime trend prediction. *Intelligent Data Analysis*, *22*(1), 117–141. https://doi.org/10.3233/IDA-163183

Agrawal, H., & Kaushal, R. (2016). Analysis of Text Mining Techniques over Public Pages of Facebook. *Proceedings of the IEEE 6th International Conference on Advanced Computing (IACC)*, 9–14. https://doi.org/10.1109/IACC.2016.12

Al-Khalisy, M. A. E., & Jehlol, H. B. (2018). Terrorist affiliations identifying through Twitter social media analysis using data mining and web mapping techniques. *Journal of Engineering and Applied Sciences*, *13*(17), 7459–7464. https://doi.org/10.36478/jeasci.2018.7459.7464

Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2020). Improving named entity recognition in noisy user-generated text with local distance neighbor feature. *Neurocomputing*, 382, 1–11. https://doi.org/10.1016/j.neucom.2019.11.072

Al-Ramahi, M., Alsmadi, I., & Davenport, J. (2020). Exploring hackers assets. *Proceedings of the 7th Symposium on Hot Topics in the Science of Security*, 1–4. https://doi.org/10.1145/3384217.3385619

AL-Saif, H., & Al-Dossari, H. (2018). Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques. *International Journal of Advanced Computer Science and Applications*, *9*(10), 377–387. https://doi.org/10.14569/IJACSA.2018.091046

Alagheband, M. R., Mashatan, A., & Zihayat, M. (2020). Time-based Gap Analysis of Cybersecurity Trends in Academic and Digital Media. *ACM Transactions on Management Information Systems*, 11(4), 1–20. https://doi.org/10.1145/3389684

Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic. *Procedia Computer Science*, *142*, 174–181. https://doi.org/10.1016/j.procs.2018.10.473

Alami, S., & Elbeqqali, O. (2015). Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts. *Proceedings of the 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 1–5. https://doi.org/10.1109/SITA.2015.7358435

Alami, S., & Elbeqqali, O. (2016). Text Mining for Suspicious Contents in Mobile Cloud Computing Environment. In M. Sabir, E and Medromi, H and Sadik (Ed.), *Advances in Ubiquitous Networking* (pp. 117–128). https://doi.org/10.1007/978-981-287-990-5_10

Alatrista-Salas, H., Morzán-Samamé, J., & Nunez-del-Prado, M. (2020). Crime Alert! Crime Typification in News Based on Text Mining. In *Lecture Notes in Networks and Systems*, Vol. 69, pp. 725–741. https://doi.org/10.1007/978-3-030-12388-8_50

Alguliyev, R. M., Aliguliyev, R. M., & Niftaliyeva, G. Y. (2018). Filtration of Terrorism-Related Texts in the E-Government Environment. *International Journal of Cyber Warfare and Terrorism*, *8*(4), 35–48. https://doi.org/10.4018/IJCWT.2018100103

Almehmadi, A., Joudaki, Z., & Jalali, R. (2017). Language usage on Twitter predicts crime rates. *Proceedings of the 10th International Conference on Security of Information and Networks - SIN '17*, 307–310. https://doi.org/10.1145/3136825.3136854

Alothman, B., & Rattadilok, P. (2017). Android botnet detection: An integrated source code mining approach. *Proceedings of the 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, 111–115. https://doi.org/10.23919/ICITST.2017.8356358

Alruily, M., Ayesh, A., & Zedan, H. (2014). Crime profiling for the Arabic language using computational linguistic techniques. *Information Processing & Management*, 50(2), 315–341. https://doi.org/10.1016/j.ipm.2013.09.001

Andleeb, S., Ahmed, R., Ahmed, Z., & Kanwal, M. (2019). Identification and Classification of Cybercrimes using Text Mining Technique. *Proceedings of the 2019 International Conference on Frontiers of Information Technology (FIT)*, 227–2275. https://doi.org/10.1109/FIT47737.2019.00050

Andriansyah, M., Purwanto, I., Subali, M., Sukowati, A. I., Samos, M., & Akbar, A. (2018). Developing Indonesian corpus of pornography using simple NLP-text mining (NTM) approach to support government anti-pornography program. *Proceedings of the Second International Conference on Informatics and Computing (ICIC)*, 1–4. https://doi.org/10.1109/IAC.2017.8280618

Angenent, M. N., Barata, A. P., & Takes, F. W. (2020). Large-scale machine learning for business sector prediction. *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 1143–1146. https://doi.org/10.1145/3341105.3374084

Anwar, T., & Abulaish, M. (2014a). A social graph based text mining framework for chat

log investigation. *Digital Investigation*, 11(4), 349–362. https://doi.org/10.1016/j.diin.2014.10.001

Anwar, T. & Abulaish, M. (2014b). Namesake alias mining on the Web and its role towards suspect tracking. *Information Sciences*, *276*, 123–145. https://doi.org/10.1016/j.ins.2014.02.050

Ariffin, N., Zainal, A., Maarof, M. A., & Kassim, M. N. (2018). A Conceptual Scheme for Ransomware Background Knowledge Construction. *Proceedings of the 2018 Cyber Resilience Conference (CRC)*, 1–4. https://doi.org/10.1109/CR.2018.8626868

Badii, A., Tiemann, M., Adderley, R., Seidler, P., Evangelio, R. H., Senst, T., … Peters, I. (2014). MOSAIC: Multimodal analytics for the protection of critical assets. *Proceedings of the 2014 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, 311–320. https://doi.org/10.0000/ieeexplore.ieee.org/7514528

Balim, C., & Gunal, E. S. (2019). Automatic Detection of Smishing Attacks by Machine Learning Methods. *Proceedings of the 1st International Informatics and Software Engineering Conference (UBMYK)*, 1–3. https://doi.org/10.1109/UBMYK48245.2019.8965429

Barbon Jr., S., Igawa, R. A., & Zarpelão, B. B. (2017). Authorship verification applied to detection of compromised accounts on online social networks. *Multimedia Tools and Applications*, 76(3), 3213–3233. https://doi.org/10.1007/s11042-016-3899-8

Bhardwaj, A., & Gupta, R. (2018). Qualitative analysis of financial statements for fraud detection. *Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 318–320. https://doi.org/10.1109/ICACCCN.2018.8748478

Basilio, M. P., Pereira, V., & Brum, G. (2019). Identification of operational demand in law enforcement agencies. *Data Technologies and Applications*, *53*(3), 333–372. https://doi.org/10.1108/DTA-12-2018-0109

Basilio, M. P., Brum, G. S., & Pereira, V. (2020). A model of policing strategy choice. *Journal of Modelling in Management*, 15(3), 849–891. https://doi.org/10.1108/JM2-10-2018-0166

Battaglia, E., Bioglio, L., & Pensa, R. G. (2020). Towards Content Sensitivity Analysis. In *Lecture Notes in Computer Science* (Vol. 12080, pp. 67–79). https://doi.org/10.1007/978-3-030-44584-3_6

Behmer, E.-J., Chandramouli, K., Garrido, V., Mühlenberg, D., Müller, D., Müller, W., … Vargas, C. (2019). Ontology Population Framework of MAGNETO for Instantiating Heterogeneous Forensic Data Modalities. In J. MacIntyre, I. Maglogiannis, L. Iliadis,

& E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (Vol. 559, pp. 520–531). https://doi.org/10.1007/978-3-030-19823-7_44

Birks, D., Coleman, A., & Jackson, D. (2020). Unsupervised identification of crime problems from police free-text data. *Crime Science*, 9(1). https://doi.org/10.1186/s40163-020-00127-4

Bisgin, H., Arslan, H., & Korkmaz, Y. (2019). Analyzing the Dabiq Magazine: The Language and the Propaganda Structure of ISIS. In R. Thomson, H. Bisgin, C. Dancy, & A. Hyder (Eds.), *Social, Cultural, and Behavioral Modeling* (Vol. 11549 LNCS, pp. 1–11). https://doi.org/10.1007/978-3-030-21741-9_1

Bisio, F., Meda, C., Zunino, R., Surlinelli, R., Scillia, E., & Ottaviano, A. (2015). Real-time monitoring of Twitter traffic by using semantic networks. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, 966–969. https://doi.org/10.1145/2808797.2809371

Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, 179, 115001. https://doi.org/10.1016/j.eswa.2021.115001

Calderon, M. H. H., Palad, E. B. B., & Tangkeko, M. S. (2020). Filipino Online Scam Data Classification using Decision Tree Algorithms. *Proceedings of the 2020 International Conference on Data Science and Its Applications (ICoDSA)*, 1–6. https://doi.org/10.1109/ICoDSA50139.2020.9212929

Cardoza, C., & Wagh, R. (2017). Text analysis framework for understanding cyber-crimes. *International Journal of Advanced and Applied Sciences*, 4(10), 58–63. https://doi.org/10.21833/ijaas.2017.010.010

Castillo-Zúñiga, I., Luna-Rosas, F. J., Rodríguez-Martínez, L. C., Muñoz-Arteaga, J., López-Veyna, J. I., & Rodríguez-Díaz, M. A. (2020). Internet Data Analysis Methodology for Cyberterrorism Vocabulary Detection, Combining Techniques of Big Data Analytics, NLP and Semantic Web. *International Journal on Semantic Web and Information Systems*, 16(1), 69–86. https://doi.org/10.4018/IJSWIS.2020010104

Cataldo, R., Galasso, R., Grassia, M. G., & Marina, M. (2017). #Theterrormood: Studying the World Mood After the Terror Attacks on Paris and Bruxelles. In N. C. Lauro, E. Amaturo, M. G. Grassia, B. Aragona, & M. Marino (Eds.), *Data Science and Social Research* (Vol. 2, pp. 185–192). https://doi.org/10.1007/978-3-319-55477-8_17

Chandra, N., Khatri, S. K., & Som, S. (2017). Anti social comment classification based on kNN algorithm. *Proceedings of the 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 348–354. https://doi.org/10.1109/ICRITO.2017.8342450

Chang, H.-C., & Wang, C.-Y. (2015). Cloud Incident Data Analytics: Change-Point Analysis and Text Visualization. *Proceedings of the 48th Hawaii International Conference on System Sciences*, 2015-March, 5320–5330. https://doi.org/10.1109/HICSS.2015.626

Chen, L.-C., Hsu, C.-L., Lo, N.-W., Yeh, K.-H., & Lin, P.-H. (2017). Fraud Analysis and Detection for Real-Time Messaging Communications on Social Networks. *IEICE Transactions on Information and Systems*, E100.D(10), 2267–2274. https://doi.org/10.1587/transinf.2016INI0003

Chen, W., Aspinall, D., Gordon, A. D., Sutton, C., & Muttik, I. (2016a). A text-mining approach to explain unwanted behaviours. *Proceedings of the 9th European Workshop on System Security - EuroSec '16*, 1–6. https://doi.org/10.1145/2905760.2905763

Chen, W., Aspinall, D., Gordon, A. D., Sutton, C., & Muttik, I. (2016b). On Robust Malware Classifiers by Verifying Unwanted Behaviours. In E. Abraham & M. Huisman (Eds.), *Integrated Formal Methods*, pp. 326–341. https://doi.org/10.1007/978-3-319-33693-0_21

Choi, Y.-J., Jeon, B.-J., & Kim, H.-W. (2021). Identification of key cyberbullies: A text mining and social network analysis approach. *Telematics and Informatics*, 56, 101504. https://doi.org/10.1016/j.tele.2020.101504

Choudhary, S. P., & Vidyarthi, M. D. (2015). A Simple Method for Detection of Metamorphic Malware using Dynamic Analysis and Text Mining. *Procedia Computer Science*, 54, 265–270. https://doi.org/10.1016/j.procs.2015.06.031

Chung, W., Liu, J., Tang, X., & Lai, V. S. K. (2018). Extracting Textual Features of Financial Social Media to Detect Cognitive Hacking. *Proceedings of the 2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 244–246. https://doi.org/10.1109/ISI.2018.8587364

Cichosz, P. (2018). A Case Study in Text Mining of Discussion Forum Posts: Classification with Bag of Words and Global Vectors. *International Journal of Applied Mathematics and Computer Science*, *28*(4), 787–801. https://doi.org/10.2478/amcs-2018-0060

Concepción-Sánchez, J. A., Molina-Gil, J., Caballero-Gil, P., & Santos-Gonzalez, I. (2018). Fuzzy Logic System for Identity Theft Detection in Social Networks. *Proceedings of the 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*, 65–70. https://doi.org/10.1109/Innovate-Data.2018.00017

Correa, J. C., García-Chitiva, M. del P., & García-Vargas, G. R. (2018). A Text Mining Approach to the Text Difficulty of Latin American Peace Agreement. *Revista Latinoamericana de Psicología*, *50*(1), 61–70. https://doi.org/10.14349/rlp.2018.v50.n1.6

Costa, E., Ferreira, R., Brito, P., Bittencourt, I. I., Holanda, O., MacHado, A., & Marinho, T.

(2012). A framework for building web mining applications in the world of blogs: A case study in product sentiment analysis. *Expert Systems with Applications*, 39(5), 4813–4834. https://doi.org/10.1016/j.eswa.2011.09.135

Das, P. & Das, A. K. (2017a). A two-stage approach of named-entity recognition for crime analysis. *Proceedings of the 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–5. https://doi.org/10.1109/ICCCNT.2017.8203949

Das, P. & Das, A. K. (2017b). An application of strength Pareto evolutionary algorithm for feature selection from crime data. *Proceedings of the 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT.2017.8203948

Das, P. & Das, A. K. (2017c). Crime analysis against women from online newspaper reports and an approach to apply it in dynamic environment. *Proceedings of the 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 312–317. https://doi.org/10.1109/ICBDACI.2017.8070855

Das, P. & Das, A. K. (2019). Graph-based clustering of extracted paraphrases for labelling crime reports. *Knowledge-Based Systems*, *179*, 55–76. https://doi.org/10.1016/j.knosys.2019.05.004

Dastjerdi, A. R., Foroghi, D., & Kiani, G. H. (2019). Detecting manager's fraud risk using text analysis: evidence from Iran. *Journal of Applied Accounting Research*, *20*(2), 154–171. https://doi.org/10.1108/JAAR-01-2018-0016

de Boer, M. H. T., Bakker, B. J., Boertjes, E., Wilmer, M., Raaijmakers, S., & van der Kleij, R. (2019). Text Mining in Cybersecurity: Exploring Threats and Opportunities. *Multimodal Technologies and Interaction*, 3(3), 62. https://doi.org/10.3390/mti3030062

de la Torre, C. J., Sánchez, D., Blanco, I., & Martín-Bautista, M. J. (2018). Text mining: Techniques, applications, and challenges. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, *26*(4), 553–582. https://doi.org/10.1142/S0218488518500265

Ding, S. H. H., Fung, B. C. M., & Debbabi, M. (2015). A Visualizable Evidence-Driven Approach for Authorship Attribution. *ACM Transactions on Information and System Security*, 17(3), 12:2-12:30. https://doi.org/10.1145/2699910

Dong, F., Yuan, S., Ou, H., & Liu, L. (2018). New Cyber Threat Discovery from Darknet Marketplaces. *Proceedings of the 2018 IEEE Conference on Big Data and Analytics (ICBDA)*, 62–67. https://doi.org/10.1109/ICBDAA.2018.8629658

Dong, W., Liao, S., & Liang, L. (2016). Financial statement fraud detection using text mining: A Systemic Functional Linguistics theory perspective. *Proceedings of the*

Pacific Asia Conference on Information Systems (PACIS 2016). Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85011024654&partnerID=40&md5=79ef7365d7cd18674dbd1fd346bb5a60

Dong, W., Xu, Y., Liao, S. S., & Feng, X. (2016). Leading effect of social media for financial fraud disclosure: A text mining based analytics. *Proceedings of the 22nd Americas Conference on Information Systems*. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84987638462&partnerID=40&md5=84c1e1ba793ddbbce05ae6f4db558ea1

Elkhawas, A. I. & Abdelbaki, N. (2018). Malware Detection using Opcode Trigram Sequence with SVM. *Proceedings of the 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 1–6. https://doi.org/10.23919/SOFTCOM.2018.8555738

Fa, Z., Geng, G.-G., Yan, Z.-W., & Lee, X.-D. (2017). A robust internet abuse detection method. *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)*, 2018-January, 1712–1715. https://doi.org/10.1109/BigData.2017.8258113

Fontanarava, J., Pasi, G., & Viviani, M. (2017). An ensemble method for the credibility assessment of user-generated content. *Proceedings of the 2017 IEEE/WIC/ACM International Conference on Web Intelligence*, 863–868. https://doi.org/10.1145/3106426.3106464

Gravanis, G., Vakali, A., Diamantaras, K., & Karadais, P. (2019). Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128, 201–213. https://doi.org/10.1016/j.eswa.2019.03.036

Giacalone, M., Cusatelli, C., Romano, A., Buondonno, A., & Santarcangelo, V. (2018). Big Data and forensics: An innovative approach for a predictable jurisprudence. *Information Sciences*, *426*, 160–170. https://doi.org/10.1016/j.ins.2017.10.036

Gil, V. D., Betancur, J. D., Puerta, I. C., Montoya, L. M., & Sepulveda, J. M. (2018). The Femicide in Colombia and Mexico: A Text Mining Analysis. *The Turkish Online Journal of Design, Art and Communication*, *2018*(SI), 170–177. https://doi.org/10.7456/1080MSE/021

Glancy, F., Biros, D. P., Liang, N., & Luse, A. (2020). Classification of malicious insiders and the association of the forms of attacks. *Journal of Criminal Psychology*, 10(3), 233–247. https://doi.org/10.1108/JCP-03-2020-0012

Gomes, T., & Ladeira, M. (2020). A new conceptual framework for enhancing legal information retrieval at the Brazilian Superior Court of Justice. *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, 26–29. https://doi.org/10.1145/3415958.3433087

Gowri, S., Anandha Mala, G. S., & Divya, G. (2014). Enhancing the digital data retrieval system using novel techniques. *Journal of Theoretical and Applied Information Technology*, *66*(2), 481–489. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84906493905&partnerID=40&md5=a9ec4a97aabddb9d470ec2136b43cc17

Hadad, T., Puzis, R., Sidik, B., Ofek, N., & Rokach, L. (2018). Application Marketplace Malware Detection by User Feedback Analysis. In P. Mori, S. Furnell, & O. Camp (Eds.), *Information Systems Security and Privacy* (Vol. 867, pp. 1–19). https://doi.org/10.1007/978-3-319-93354-2_1

Hadad, T., Sidik, B., Ofek, N., Puzis, R., & Rokach, L. (2017). User Feedback Analysis for Mobile Malware Detection. *Proceedings of the 3rd International Conference on Information Systems Security and Privacy*, 83–94. https://doi.org/10.5220/0006131200830094

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, 128, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Halouzka, K., & Burita, L. (2019). Cyber Security Strategic Documents Analysis. *Proceedings of the 2019 International Conference on Military Technologies (ICMT)*, 1–6. https://doi.org/10.1109/MILTECHS.2019.8870088

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Retrieved from http://www.amazon.co.uk/Data-Mining-Concepts-Techniques-Management/dp/0123814790

Hao, J., & Dai, H. (2016). Social media content and sentiment analysis on consumer security breaches. *Journal of Financial Crime*, 23(4), 855–869. https://doi.org/10.1108/JFC-01-2016-0001

Henseler, H., & Hyde, J. (2019). Technology assisted analysis of timeline and connections in digital forensic investigations. *CEUR Workshop Proceedings*, 2484, 32–37. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076051990&partnerID=40&md5=f2e474c504642f0d296c932a51e9f399

Hicks, C., Beebe, N. L., & Haliscak, B. (2016). Extending web mining to digital forensics text mining. *Proceedings of the 22nd Americas Conference on Information Systems*. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84987624224&partnerID=40&md5=aa7e54ff003d8549a16fb8cc13c41e5c

Hernandez-Castro, J., & Roberts, D. L. (2015). Automatic detection of potentially illegal online sales of elephant ivory via data mining. *PeerJ Computer Science*, 1(7).

https://doi.org/10.7717/peerj-cs.10

Hou, J., Li, X., Yao, H., Sun, H., Mai, T., & Zhu, R. (2020). BERT-Based Chinese Relation Extraction for Public Security. *IEEE Access*, 8, 132367–132375. https://doi.org/10.1109/ACCESS.2020.3002863

Hultgren, M., Whitney, J., Jennex, M. E., & Elkins, A. (2018). A Knowledge Management Approach to Identify Victims of Human Sex Trafficking. *Communications of the Association for Information Systems*, *42*(1), 602–620. https://doi.org/10.17705/1CAIS.04223

Husari, G., Al-Shaer, E., Ahmed, M., Chu, B., & Niu, X. (2017). TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources. *Proceedings of the 33rd Annual Computer Security Applications Conference*, Part F1325, 103–115. https://doi.org/10.1145/3134600.3134646

Husari, G., Niu, X., Chu, B., & Al-Shaer, E. (2018). Using Entropy and Mutual Information to Extract Threat Actions from Cyber Threat Intelligence. *Proceedings of the 2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1–6. https://doi.org/10.1109/ISI.2018.8587343

Iftikhar, A., Jaffry, S. W. U. Q., & Malik, M. K. (2019). Information Mining From Criminal Judgments of Lahore High Court. *IEEE Access*, *7*, 59539–59547. https://doi.org/10.1109/ACCESS.2019.2915352

Ishara Amali, H. M. A., & Jayalal, S. (2020). Classification of Cyberbullying Sinhala Language Comments on Social Media. *Proceedings of the 2020 Moratuwa Engineering Research Conference (MERCon)*, 266–271. https://doi.org/10.1109/MERCon50084.2020.9185209

Jackson, P., & Moulinier, I. (2002). *Natural language processing for online applications: text retrieval, extraction, and categorization*. John Benjamins Publishing Company.

Johnston, A. H., & Weiss, G. M. (2017). Identifying Sunni extremist propaganda with deep learning. *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–6. https://doi.org/10.1109/SSCI.2017.8280944

Kabwe, F., & Phiri, J. (2020). Identity attributes metric modelling based on mathematical distance metrics models. *International Journal of Advanced Computer Science and Applications*, 11(7), 450–464. https://doi.org/10.14569/IJACSA.2020.0110759

Kakavand, M., Mustapha, A., Tan, Z., Yazdani, S. F., & Arulsamy, L. (2019). O-ADPI: Online Adaptive Deep-Packet Inspector Using Mahalanobis Distance Map for Web Service Attacks Classification. *IEEE Access*, 7, 167141–167156. https://doi.org/10.1109/ACCESS.2019.2953791

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer London. https://doi.org/10.1007/978-1-84628-754-1

Karystianis, G., Adily, A., Schofield, P., Knight, L., Galdon, C., Greenberg, D., … Butler, T. (2018). Automatic Extraction of Mental Health Disorders From Domestic Violence Police Narratives: Text Mining Study. *Journal of Medical Internet Research*, *20*(9), e11548. https://doi.org/10.2196/11548

Karystianis, G., Adily, A., Schofield, P. W., Greenberg, D., Jorm, L., Nenadic, G., & Butler, T. (2019). Automated Analysis of Domestic Violence Police Reports to Explore Abuse Types and Victim Injuries: Text Mining Study. *Journal of Medical Internet Research*, *21*(3), e13067. https://doi.org/10.2196/13067

Karystianis, G., Simpson, A., Adily, A., Schofield, P., Greenberg, D., Wand, H., Nenadic, G., & Butler, T. (2020). Prevalence of Mental Illnesses in Domestic Violence Police Records: Text Mining Study. *Journal of Medical Internet Research*, 22(12), e23725. https://doi.org/10.2196/23725

Kavita, Mahani, P., & Ruhil, N. (2016). Web data mining: A perspective of research issues and challenges. *3rd International Conference on Computing for Sustainable Global Development*, 3235–3238. Retrieved from http://ieeexplore-ieee-org.ez9.periodicos.capes.gov.br/stamp/stamp.jsp?tp=&arnumber=7724863&isnumber=7724213

Kaur, H., Choudhury, T., Singh, T. P., & Shamoon, M. (2019). Crime Analysis using Text Mining. *Proceedings of the 2019 International Conference on Contemporary Computing and Informatics (IC3I)*, 283–288. https://doi.org/10.1109/IC3I46837.2019.9055606

Kim, Y., Cho, S., Han, S., & You, I. (2018). A software classification scheme using binary-level characteristics for efficient software filtering. *Soft Computing*, *22*(2), 595–606. https://doi.org/10.1007/s00500-016-2357-x

Kitchenham, B. & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. In *EBSE Technical Report EBSE-2007-01*.

Kuang, D., Brantingham, P. J., & Bertozzi, A. L. (2017). Crime topic modeling. *Crime Science*, *6*(1), 12. https://doi.org/10.1186/s40163-017-0074-0

Kumar, A. S. & Palanisamy, N. (2008). Challenges for Web Mining. *Proceedings of the 2008 International Conference on Computing, Communication and Networking, ICCCN 2008*. https://doi.org/10.1109/ICCCNET.2008.4787768

Kumar, G. R., Mangathayaru, N., & Narasimha, G. (2015). An approach for Intrusion Detection using Text Mining Techniques. *Proceedings of the International Conference on Engineering & MIS 2015*, 24-26-Sept, 1–6.

https://doi.org/10.1145/2832987.2833076

Lal, S., Tiwari, L., Ranjan, R., Verma, A., Sardana, N., & Mourya, R. (2020). Analysis and Classification of Crime Tweets. *Procedia Computer Science*, 167, 1911–1919. https://doi.org/10.1016/j.procs.2020.03.211

Lee, P. S., Owda, M., & Crockett, K. (2018). Novel Methods for Resolving False Positives during the Detection of Fraudulent Activities on Stock Market Financial Discussion Boards. *International Journal of Advanced Computer Science and Applications*, *9*(1), 1–10. https://doi.org/10.14569/IJACSA.2018.090101

Lee, T.-H., Sung, W.-K., & Kim, H.-W. (2016). A text mining approach to the analysis of information security awareness: Korea, United States, and China. *Proceedings of the Pacific Asia Conference on Information Systems*. Retrieved from https://aisel.aisnet.org/pacis2016/69

Lekea, I., & Karampelas, P. (2017). Are We Really That Close Together? Tracing and Discussing Similarities and Differences between Greek Terrorist Groups Using Cluster Analysis. *Proceedings of the 2017 European Intelligence and Security Informatics Conference (EISIC)*, 159–162. https://doi.org/10.1109/EISIC.2017.33

Li, L., Xiao, W., Dai, C., Tong, H., & Song, Z. (2014). Mining the Association of Multiple Virtual Identities Based on Multi-Agent Interaction. In H. Wang & M. A. Sharaf (Eds.), *Databases Theory and Applications* (Vol. 8506 LNCS, pp. 172–179). https://doi.org/10.1007/978-3-319-08608-8_15

Li, W., Chen, H., & Nunamaker, J. F. (2016). Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System. *Journal of Management Information Systems*, 33(4), 1059–1086. https://doi.org/10.1080/07421222.2016.1267528

Li, X., Xu, W., & Tian, X. (2014). How to protect investors? A GA-based DWD approach for financial statement fraud detection. *Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3548–3554. https://doi.org/10.1109/SMC.2014.6974480

Liang, N., Biros, D.P., & Luse, A. (2016). An Empirical Validation of Malicious Insider Characteristics. *Journal of Management Information Systems*, 33(2), 361–392. https://doi.org/10.1080/07421222.2016.1205925

Liang, N. & Biros, D. (2016). Validating Common Characteristics of Malicious Insiders: Proof of Concept Study. *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS)*, 3716–3726. https://doi.org/10.1109/HICSS.2016.463

Lyu, Y., Chow, J. C.-C., & Hwang, J.-J. (2020). Exploring public attitudes of child abuse in mainland China: A sentiment analysis of China's social media Weibo. *Children and*

Youth Services Review, 116. https://doi.org/10.1016/j.childyouth.2020.105250

Ma, R., Basumallik, S., & Eftekharnejad, S. (2020). A PMU-Based Data-Driven Approach for Classifying Power System Events Considering Cyberattacks. *IEEE Systems Journal*, 14(3), 3558–3569. https://doi.org/10.1109/JSYST.2019.2963546

Maktabar, M., Zainal, A., Maarof, M. A., & Kassim, M. N. (2018). Content Based Fraudulent Website Detection Using Supervised Machine Learning Techniques. In A. Abraham, P. K. Muhuri, A. K. Muda, & N. Gandhi (Eds.), *Hybrid Intelligent Systems*, Vol. 734, pp. 294–304. https://doi.org/10.1007/978-3-319-76351-4_30

Malim, N. H. A. H., Sagadevan, S., & Ridzuwan, N. I. (2019). Criminality recognition using machine learning on Malay language tweets. *Pertanika Journal of Science and Technology*, 27(4), 1803–1820. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074112842&partnerID=40&md5=26ec0607eb6aa010d3f52d95a1db2777

Mansour, S. (2018). Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining. *Procedia Computer Science*, *140*, 95–103. https://doi.org/10.1016/j.procs.2018.10.297

Margono, H., Yi, X., & Raikundalia, G. K. (2014). Mining Indonesian Cyber Bullying Patterns in Social Networks. *Proceedings of the 37th Australasian Computer Science Conference*, *147*, 115–124. Retrieved from https://dl.acm.org/doi/abs/10.5555/2667473.2667487

Marivate, V., & Moiloa, P. (2016). Catching crime: Detection of public safety incidents using social media. *Proceedings of the 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 1–5. https://doi.org/10.1109/RoboMech.2016.7813140

Martin, A., Calleja, A., Menendez, H. D., Tapiador, J., & Camacho, D. (2016). ADROIT: Android malware detection using meta-information. *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. https://doi.org/10.1109/SSCI.2016.7849904

Martinelli, F., Marulli, F., & Mercaldo, F. (2017). Evaluating Convolutional Neural Network for Effective Mobile Malware Detection. *Procedia Computer Science*, 112, 2372–2381. https://doi.org/10.1016/j.procs.2017.08.216

Marulli, F., & Mercaldo, F. (2017). Let's Gossip: Exploring Malware Zero-Day Time Windows by Social Network Analysis. *Proceedings of the 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 704–709. https://doi.org/10.1109/WAINA.2017.114

Miah, M.W.R., Yearwood, J., & Kulkarni, S. (2015). Constructing an inter-post similarity measure to differentiate the psychological stages in offensive chats. *Journal of the*

Association for Information Science and Technology, 66(5), 1065–1081. https://doi.org/10.1002/asi.23247

Mine, T., Hirokawa, S., & Suzuki, T. (2019). Does Crime Activity Report Reveal Regional Characteristics? In S. Lee, R. Ismail, & H. Choo (Eds.), *Advances in Intelligent Systems and Computing*, Vol. 935, pp. 582–598. https://doi.org/10.1007/978-3-030-19063-7_46

Miranda, E., Aryuni, M., Fernando, Y., & Kibtiah, T. M. (2020). A study of radicalism contents detection in Twitter: Insights from support vector machine technique. *Proceedings of the 2020 International Conference on Information Management and Technology (ICIMTech 2020)*, 549–554. https://doi.org/10.1109/ICIMTech50083.2020.9211229

Mishra, S., Shukla, P. K., & Agarwal, R. (2020). Location wise opinion mining of real time Twitter data using Hadoop to reduce cyber crimes. *Proceedings of the 2nd International Conference on Data, Engineering and Applications (IDEA 2020)*. https://doi.org/10.1109/IDEA49133.2020.9170700

Mohasseb, A., Aziz, B., Jung, J., & Lee, J. (2019). Predicting CyberSecurity Incidents using Machine Learning Algorithms: A Case Study of Korean SMEs. *Proceedings of the 5th International Conference on Information Systems Security and Privacy*, 230–237. https://doi.org/10.5220/0007309302300237

Monish, H., & Pandey, A. C. (2020). A comparative assessment of data mining algorithms to predict fraudulent firms. *Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering*, 117–122. https://doi.org/10.1109/Confluence47617.2020.9057968

Mukherjee, S., & Sarkar, K. (2020). Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas. *Proceedings of the 2020 IEEE Calcutta Conference (CALCON 2020)*, 444–450. https://doi.org/10.1109/CALCON49167.2020.9106554

Nedeljkovic, S., Nikolic, V., Cabarkapa, M., Misic, J., & Randelovic, D. (2019). An Advanced Quick-Answering System Intended for the e-Government Service in the Republic of Serbia. *Acta Polytechnica Hungarica*, *16*(4), 153–174. https://doi.org/10.12700/APH.16.4.2019.4.8

Netsuwan, T., & Kesorn, K. (2017). Unify framework for crime data summarization using RSS feed service. *Walailak Journal of Science and Technology*, 14(10), 769–781.

Neumann, M., & Sartor, N. (2016). A Semantic Network Analysis of Laundering Drug Money. *Journal of Tax Administration*, 2(1), 73–94.

Niekerk, B. van, Ramluckan, T., & Duvenage, P. (2019). An analysis of selected cyber intelligence texts. *Proceedings of the 18th European Conference on Cyber Warfare and*

Security, 2019-July, 551–559. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85070014329&partnerID=40&md5=c12a9cf42b009feec6f7b2eaf4da79fd

Nikolić, V., Markoski, B., Ivkovic, M., Kuk, K., & Djikanovic, P. (2015). Information retrieval for unstructured text documents in Serbian into the crime domain. *Proceedings of the 16th IEEE International Symposium on Computational Intelligence and Informatics*, 267–271. https://doi.org/10.1109/CINTI.2015.7382934

Ning, Y., Muthiah, S., Rangwala, H., & Ramakrishnan, N. (2016). Modeling Precursors for Event Forecasting via Nested Multi-Instance Learning. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, San Francisco, 1095–1104. https://doi.org/10.1145/2939672.2939802

Nguyen, A., Hoang, Q., Nguyen, H., Nguyen, D., & Tran, T. (2017). Evaluating marijuana-related tweets on Twitter. *Proceedings of the IEEE 7th Annual Computing and Communication Workshop and Conference*, 1–7. https://doi.org/10.1109/CCWC.2017.7868364

Noel, G. E. & Peterson, G. L. (2014). Applicability of Latent Dirichlet Allocation to multi-disk search. *Digital Investigation*, *11*(1), 43–56. https://doi.org/10.1016/j.diin.2014.02.001

Noviantho, Isa, S. M., & Ashianti, L. (2017). Cyberbullying classification using text mining. *Proceedings of the 1st International Conference on Informatics and Computational Sciences (ICICoS)*, 241–246. https://doi.org/10.1109/ICICOS.2017.8276369

Nwafor, E., Chowdhary, P., & Chandra, A. (2016). A Policy-Driven Framework for Document Classification and Enterprise Security. *Proceedings of the 2016 International IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom*, 949–953. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0149

Overbeck, M. (2015). Observers turning into participants: Shifting perspectives on religion and armed conflict in western news coverage. *Tocqueville Review*, 36(2), 95–124. Retrieved from https://muse.jhu.edu/article/610761

Owda, M., Lee, P. S., & Crockett, K. (2017). Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using information extraction. *Proceedings of the 2017 Intelligent Systems Conference (IntelliSys)*, 1078–1082. https://doi.org/10.1109/IntelliSys.2017.8324262

Öztürk, N. & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, *35*(1), 136–147.

https://doi.org/10.1016/j.tele.2017.10.006

Palad, E. B. B., Tangkeko, M. S., Magpantay, L. A. K., & Sipin, G. L. (2019). Document Classification of Filipino Online Scam Incident Text using Data Mining Techniques. *Proceedings of the 19th International Symposium on Communications and Information Technologies*, ISCIT 2019, 232–237. https://doi.org/10.1109/ISCIT.2019.8905242

Palad, E. B. B., Burden, M. J. F., Torre, C. R. Dela, & C. Uy, R. B. (2020). Performance evaluation of decision tree classification algorithms using fraud datasets. *Bulletin of Electrical Engineering and Informatics*, 9(6), 2518–2525. https://doi.org/10.11591/eei.v9i6.2630

Parapar, J., Losada, D. E., & Barreiro, Á. (2014). Combining psycho-linguistic, content-based and chat-based features to detect predation in chatrooms. *Journal of Universal Computer Science*, *20*(2), 213–239. https://doi.org/10.3217/jucs-020-02-0213

Park, J., Park, C., Kim, J., Cho, M., & Park, S. (2019). ADC: Advanced document clustering using contextualized representations. *Expert Systems with Applications*, 137, 157–166. https://doi.org/10.1016/j.eswa.2019.06.068

Percy, I., Balinsky, A., Balinsky, H., & Simske, S. (2018). Text Mining and Recommender Systems for Predictive Policing. *Proceedings of the ACM Symposium on Document Engineering 2018*, 1–4. https://doi.org/10.1145/3209280.3229112

Petrovskiy, M. & Chikunov, M. (2019). Online Extremism Discovering through Social Network Structure Analysis. *Proceedings of the IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, 243–249. https://doi.org/10.1109/INFOCT.2019.8711254

Pina-Sánchez, J., Roberts, J. V, & Sferopoulos, D. (2019a). Does the Crown Court Discriminate Against Muslim-named Offenders? a Novel Investigation Based on Text Mining Techniques. *The British Journal of Criminology*, *59*(3), 718–736. https://doi.org/10.1093/bjc/azy062

Pina-Sánchez, J., Grech, D., Brunton-Smith, I., & Sferopoulos, D. (2019b). Exploring the origin of sentencing disparities in the Crown Court: Using text mining techniques to differentiate between court and judge disparities. *Social Science Research*, 84, 102343. https://doi.org/10.1016/j.ssresearch.2019.102343

Pires, M., & Georgieva, P. (2020). An Intelligent Tool for Detection of Phishing Messages. In *Advances in Intelligent Systems and Computing* (Vol. 942, pp. 116–125). https://doi.org/10.1007/978-3-030-17065-3_12

Po, L. & Rollo, F. (2018). Building an Urban Theft Map by Analyzing Newspaper Crime Reports. *Proceedings of the 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, 13–18.

https://doi.org/10.1109/SMAP.2018.8501866

Qazi, N. & Wong, B. L. W. (2019). An interactive human centered data science approach towards crime pattern analysis. *Information Processing & Management*, *56*(6). https://doi.org/10.1016/j.ipm.2019.102066

Rabuzin, K., & Modrušan, N. (2019). Prediction of public procurement corruption indices using machine learning methods. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 3, 333–340. https://doi.org/10.5220/0008353603330340

Rehman, Z. U., Abbas, S., Khan, M. A., Mustafa, G., Fayyaz, H., Hanif, M., & Saeed, M. A. (2020). Understanding the language of ISIS: An empirical approach to detect radical content on Twitter using machine learning. *Computers, Materials and Continua*, 66(2), 1075–1090. https://doi.org/10.32604/cmc.2020.012770

Ristea, A., Kurland, J., Resch, B., Leitner, M., & Langford, C. (2018). Estimating the Spatial Distribution of Crime Events around a Football Stadium from Georeferenced Tweets. *ISPRS International Journal of Geo-Information*, *7*(2), 43. https://doi.org/10.3390/ijgi7020043

Robinson, P. H., & Dubber, M. D. (2007). The American Model Penal Code: A Brief Overview. *New Criminal Law Review*, 10(3), 319-341. https://doi.org/10.1525/nclr.2007.10.3.319

Roopa, V. & Induja, K. (2019). Customized Visualization of Email Using Sentimental and Impact Analysis in R. In M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, T. Ören, & R. Kashyap (Eds.), *Advances in Computing and Data Sciences*, Vol. 1046, pp. 144–154. https://doi.org/10.1007/978-981-13-9942-8_14

Ruano-Ordás, D., Fdez-Riverola, F., & Méndez, J. R. (2018). Concept drift in e-mail datasets: An empirical study with practical implications. *Information Sciences*, *428*, 120–135. https://doi.org/10.1016/j.ins.2017.10.049

Saha, P., Bose, I., & Mahanti, A. (2016). A knowledge based scheme for risk assessment in loan processing by banks. *Decision Support Systems*, 84, 78–88. https://doi.org/10.1016/j.dss.2016.02.002

Sahu, S., Agrawal, S., & Baraskar, R. (2019). The effect of best first search optimization on credit card fraudulent transaction detection. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), 1939–1946. https://doi.org/10.35940/ijitee.L2894.1081219

Saini, J. K., & Bansal, D. (2019). A Comparative Study and Automated Detection of Illegal Weapon Procurement over Dark Web. *Cybernetics and Systems*, *50*(5), 405–416. https://doi.org/10.1080/01969722.2018.1553591

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

Sameera, K., & Vishwakarma, P. (2019). Cybercrime: To Detect Suspected User's Chat Using Text Mining. In S. C. Satapathy & A. Joshi (Eds.), *Information and Communication Technology for Intelligent Systems* (Vol. 106, pp. 381–390). https://doi.org/10.1007/978-981-13-1742-2_37

Samtani, S., Chinn, R., Chen, H., & Nunamaker, J. F. (2017). Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*, *34*(4), 1023–1053. https://doi.org/10.1080/07421222.2017.1394049

Samtani, S., Yu, S., Zhu, H., Patton, M., Matherly, J., & Chen, H. (2018). Identifying SCADA Systems and Their Vulnerabilities on the Internet of Things: A Text-Mining Approach. *IEEE Intelligent Systems*, *33*(2), 63–73. https://doi.org/10.1109/MIS.2018.111145022

Samtani, S., Zhu, H., & Chen, H. (2020). Proactively Identifying Emerging Hacker Threats from the Dark Web. *ACM Transactions on Privacy and Security*, 23(4), 1–33. https://doi.org/10.1145/3409289

Sankar, K., Jackovich, J., & Richards, R. (2020). *Applied AI and Natural Language Processing Workshop*. Packt Publishing Ltd.

Sarwar, S., Qayyum, Z. U., Safyan, M., Iqbal, M., & Mahmood, Y. (2019). Graphs Resemblance based Software Birthmarks through Data Mining for Piracy Control. *Programming and Computer Software*, 45(8), 581–589. https://doi.org/10.1134/S0361768819080152

Savaliya, B. R., & Philip, C. G. (2017). Email fraud detection by identifying email sender. *Proceedings of the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 1420–1422. https://doi.org/10.1109/ICECDS.2017.8389678

Savaş, S. & Topaloğlu, N. (2019). Data analysis through social media according to the classified crime. *Turkish Journal of Electrical Engineering & Computer Sciences*, *27*(1), 407–420. https://doi.org/10.3906/elk-1712-17

Seidler, P., Adderley, R., Badii, A., & Raffaelli, M. (2014). MOSAIC: Criminal network analysis for multi-modal surveillance and decision support. *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 257–260. https://doi.org/10.1109/ASONAM.2014.6921593

Sharef, N. M., & Martin, T. (2015). Evolving fuzzy grammar for crime texts categorization. *Applied Soft Computing*, 28, 175–187. https://doi.org/10.1016/j.asoc.2014.11.038

Sharmin, S., & Zaman, Z. (2017). Spam Detection in Social Media Employing Machine Learning Tool for Text Mining. *Proceedings of the 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 137–142. https://doi.org/10.1109/SITIS.2017.32

Siering, M., Muntermann, J., & Grčar, M. (2021). Design principles for robust fraud detection: The case of stock market manipulations. *Journal of the Association for Information Systems*, 22(1), 156–178. https://doi.org/10.17705/1jais.00657

Silomon, J. A. M. & Roeling, M. P. (2018). Assessing Opinions on Software as a Weapon in the Context of (Inter)national Security. In M. L. Gavrilova, C. J. K. Tan, & A. Sourin (Eds.), *Transactions on Computational Science XXXII* (Vol. 10830 LNCS, pp. 43–56). https://doi.org/10.1007/978-3-662-56672-5_4

Slamet, C., Krismunandar, A., Maylawati, D. S., Jumadi, Amin, A. S., & Ramdhani, M. A. (2020). Deep learning approach for bullying classification on Twitter social media with Indonesian language. *Proceedings of the 6th International Conference on Wireless and Telematics (ICWT 2020), 1-5*. https://doi.org/10.1109/ICWT50448.2020.9243653

Sonowal, G. & Kuppusamy, K. S. (2018). SmiDCA: An Anti-Smishing Model with Machine Learning Approach. *The Computer Journal*, *61*(8), 1143–1157. https://doi.org/10.1093/comjnl/bxy039

Spitters, M., Klaver, F., Koot, G., & Staalduinen, M. van. (2015). Authorship Analysis on Dark Marketplace Forums. *Proceedings of the 2015 European Intelligence and Security Informatics Conference*, 1–8. https://doi.org/10.1109/EISIC.2015.47

Suarez-Tangil, G., Tapiador, J. E., Peris-Lopez, P., & Blasco, J. (2014). Dendroid: A text mining approach to analyzing and classifying code structures in Android malware families. *Expert Systems with Applications*, 41(4), 1104–1117. https://doi.org/10.1016/j.eswa.2013.07.106

Subhan, M., Sudarsono, A., & Barakbah, A. (2017). Preprocessing of radicalism dataset to predict radical content in Indonesia. *Proceedings of the 2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, 270–275. https://doi.org/10.1109/KCIC.2017.8228598

Sundarkumar, G. G., Ravi, V., Nwogu, I., & Govindaraju, V. (2015). Malware detection via API calls, topic models and machine learning. *Proceedings of the 2015 IEEE International Conference on Automation Science and Engineering (CASE)*, 1212–1217. https://doi.org/10.1109/CoASE.2015.7294263

Sudha, T. S., & Rupa, C. (2019). Analysis and Evaluation of Integrated Cyber Crime

Offences. *Proceedings of the 2019 Innovations in Power and Advanced Computing Technologies*, i-PACT 2019. https://doi.org/10.1109/i-PACT44901.2019.8960187

Tajuddin, T., Manaf, A. A., Fatimah Awang, N., Muhamat Dawam, S. R., Rasidah Ali, N., & Amat, R. (2019). Crime Suspect Profiling (CSP) for Forensic Investigation on Smartphone. *Proceedings of the 4th International Conference and Workshops on Recent Advances and Innovations in Engineering: Thriving Technologies*. https://doi.org/10.1109/ICRAIE47735.2019.9037772

Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414–418. https://doi.org/10.14569/IJACSA.2016.071153

Tayal, K., & Ravi, V. (2016). Particle Swarm Optimization Trained Class Association Rule Mining. *Proceedings of the International Conference on Informatics and Analytics*, 25-26-August, 1–8. https://doi.org/10.1145/2980258.2980291

Thao, T. P., Yamada, A., Murakami, K., Urakawa, J., Sawaya, Y., & Kubota, A. (2017). Classification of Landing and Distribution Domains Using Whois' Text Mining. *Proceedings of the 2017 IEEE Trustcom/BigDataSE/ICESS*, 1–8. https://doi.org/10.1109/Trustcom/BigDataSE/ICESS.2017.213

Tran, Y. H., & Tran, Q. N. (2018). Estimating Public Opinion in Social Media Content Using Aspect-Based Opinion Mining. In J. Hu, I. Khalil, Z. Tari, & S. Wen (Eds.), *Mobile Networks and Management*. https://doi.org/10.1007/978-3-319-90775-8

Trovati, M., Hill, R., & Bessis, N. (2015). A Non-genuine Message Detection Method Based on Unstructured Datasets. *Proceedings of the 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 597–600. https://doi.org/10.1109/3PGCIC.2015.108

Tutun, S., Khasawneh, M. T., & Zhuang, J. (2017). New framework that uses patterns and relations to understand terrorist behaviors. *Expert Systems with Applications*, 78, 358–375. https://doi.org/10.1016/j.eswa.2017.02.029

Venčkauskas, A., Karpavičius, A., Damaševičius, R., Marcinkevičius, R., Kapočiūtė-Dzikienė, J., & Napoli, C. (2017). Open Class Authorship Attribution of Lithuanian Internet Comments using One-Class Classifier. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, *11*, 373–382. https://doi.org/10.15439/2017F461

Ventirozos, F. K., Varlamis, I., & Tsatsaronis, G. (2018). Detecting Aggressive Behavior in Discussion Threads Using Text Mining. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (*Vol. 10762 LNCS*, pp. 420–431). https://doi.org/10.1007/978-3-319-77116-8_31

Vidyarthi, D., Choudhary, S. P., Rakshit, S., & Kumar, C. R. S. (2017). Malware Detection by Static Checking and Dynamic Analysis of Executables. *International Journal of Information Security and Privacy*, 11(3), 29–41. https://doi.org/10.4018/IJISP.2017070103

Wajid, F., & Samet, H. (2016). CrimeStand: Spatial Tracking of Criminal Activity. *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1–4. https://doi.org/10.1145/2996913.2997006

Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y. (2020). Multi-modal cyberbullying detection on social networks. *Proceedings of the International Joint Conference on Neural Networks*, 1-8. https://doi.org/10.1109/IJCNN48605.2020.9206663

Wei, K., Lin, Y.-R., & Yan, M. (2020). Examining Protest as An Intervention to Reduce Online Prejudice: A Case Study of Prejudice Against Immigrants. *Proceedings of The Web Conference 2020*, 2443–2454. https://doi.org/10.1145/3366423.3380307

Xia, Y., Cai, T., & Zhong, H. (2019). Effect of judges' gender on rape sentencing: A data mining approach to analyze judgment documents. *China Review*, *19*(2), 125–149. Retrieved from https://muse.jhu.edu/article/726726#info_wrap

Xianghui, Z., Yong, P., Zan, Z., Yi, J., & Yuangang, Y. (2015). Research on Parallel Vulnerabilities Discovery Based on Open Source Database and Text Mining. *Proceedings of the 2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, 327–332. https://doi.org/10.1109/IIH-MSP.2015.84

Xylogiannopoulos, K., Karampelas, P., & Alhajj, R. (2017). Text Mining in Unclean, Noisy or Scrambled Datasets for Digital Forensics Analytics. *Proceedings of the 2017 European Intelligence and Security Informatics Conference (EISIC)*, 2017-Janua, 76–83. https://doi.org/10.1109/EISIC.2017.19

Yang, Y., Manoharan, M., & Barber, K. S. (2014). Modelling and Analysis of Identity Threat Behaviors through Text Mining of Identity Theft Stories. *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, 184–191. https://doi.org/10.1109/JISIC.2014.35

Yang, B., Jiang, J., & Li, N. (2016). Towards Discovering Covert Communication Through Email Spam. In Z. Shi, S. Vadera, & G. Li (Eds.), *Intelligent Information Processing VIII* (Vol. 486, pp. 191–201). https://doi.org/10.1007/978-3-319-48390-0_20

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

Zahra, K., Azam, F., Butt, W. H., & Ilyas, F. (2018). A Framework for User Characterization

based on Tweets Using Machine Learning Algorithms. *Proceedings of the II International Conference on Network, Communication and Computing*, 11–16. https://doi.org/10.1145/3301326.3301373

Zainal, K., Jali, M. Z., & Hasan, A. B. (2018). Comparative Analysis of Danger Theory Variants in Measuring Risk Level for Text Spam Messages. In *Advances in Intelligent Systems and Computing* (Vol. 753, pp. 133–152). https://doi.org/10.1007/978-3-319-78753-4_11

Zainol, Z., Jaymes, M. T. H., & Nohuddin, P. N. E. (2018). VisualUrText: A Text Analytics Tool for Unstructured Textual Data. *Journal of Physics: Conference Series*, *1018*, 012011. https://doi.org/10.1088/1742-6596/1018/1/012011

Zaki, M., & Theodoulidis, B. (2014). Analyzing stock market fraud cases using a linguistics-based text mining approach. In A. García-Crespo, J. M. G. Berbís, M. Radzimski, J. L. S. Cervantes, S. Coppens, K. Hammar, … M. Vander Sande (Eds.), *Joint Proceedings of the Second International Workshop on Semantic Web Enterprise Adoption and Best Practice and Second International Workshop on Finance and Economics on the Semantic Web*, Vol. 1240, pp. 63–74. Anissaras: CEUR-WS.

Zareapoor, M., & Seeja, K. R. R. (2015). Text Mining for Phishing E-mail Detection. In Intelligent Computing, Communication and Devices (Vol. 308 AISC, pp. 65–71). https://doi.org/10.1007/978-81-322-2012-1_8

Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on Distributed Computing and Networking (ICDCN '16)*, 1–6. https://doi.org/10.1145/2833312.2849567

Zhao, R., & Mao, K. (2017). Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. *IEEE Transactions on Affective Computing*, 8(3), 328–339. https://doi.org/10.1109/TAFFC.2016.2531682

Appendix A – Complementary Tables

Table A1. Complete list of journals with articles about text mining in public security published between 2014 and 2021.

Journal	Count	Journal	Count
Procedia Computer Science	5	Intelligent Data Analysis	1
Expert Systems with Applications	4	International Journal of Advanced and Applied Sciences	1
IEEE Access	3	International Journal of Applied Mathematics and Computer Science	1
Information Sciences	3	International Journal of Cyber Warfare and Terrorism	1
International Journal of Advanced Computer Science and Applications	3	International Journal of Information Security and Privacy	1
Journal of Management Information Systems	3	International Journal of Innovative Technology and Exploring Engineering	1
Journal of Medical Internet Research	3	International Journal on Semantic Web and Information Systems	1
Crime Science	2	Journal of Applied Accounting Research	1
Digital Investigation	2	Journal of Criminal Psychology	1
Information Processing & Management	2	Journal of Engineering and Applied Sciences	1
Knowledge-Based Systems	2	Journal of Financial Crime	1
Telematics and Informatics	2	Journal of Modelling in Management	1
ACM Transactions on Information and System Security	1	Journal of Tax Administration	1
ACM Transactions on Management Information Systems	1	Journal of Theoretical and Applied Information Technology	1
ACM Transactions on Privacy and Security	1	Journal of Universal Computer Science	1
Acta Polytechnica Hungarica	1	Journal of the Association for Information Science and Technology	1
Applied Soft Computing	1	Journal of the Association for Information Systems	1
Bulletin of Electrical Engineering and Informatics	1	Multimedia Tools and Applications	1
Children and Youth Services Review	1	Multimodal Technologies and Interaction	1
China Review	1	Neurocomputing	1
Communications of the Association for Information Systems	1	PeerJ Computer Science	1
Computers & Security	1	Pertanika Journal of Science and Technology	1
Computers, Materials and Continua	1	Programming and Computer Software	1
Cybernetics and Systems	1	Revista Latinoamericana de Psicología	1
Data Technologies and Applications	1	Social Science Research	1

Decision Support Systems	1	Soft Computing	1
IEEE Intelligent Systems	1	The British Journal of Criminology	1
IEEE Systems Journal	1	The Computer Journal	1
IEEE Transactions on Affective Computing	1	The Turkish Online Journal of Design, Art and Communication	1
IEICE Transactions on Information and Systems	1	Tocqueville Review	1
ISPRS International Journal of Geo-Information	1	Turkish Journal of Electrical Engineering & Computer Sciences	1
Information Systems Security and Privacy	1	Walailak Journal of Science and Technology	1
Total 1	54	Total 2	32
		Total 1 + Total 2	**86**

Table A2. Complete list of conferences with works about text mining in public security published between 2014 and 2021.

Conference	**Count**	**Conference**	**Count**
2016 Pacific Asia Conference on Information Systems	2	2018 International Conference on Advances in Computing, Communication Control and Networking	1
2017 European Intelligence and Security Informatics Conference	2	2019 Innovations in Power and Advanced Computing Technologies	1
2018 IEEE International Conference on Intelligence and Security Informatics	2	2019 International Conference on contemporary Computing and Informatics	1
22nd Americas Conference on Information Systems	2	2019 International Conference on Frontiers of Information Technology	1
8th International Conference on Computing, Communication and Networking Technologies	2	2019 International Conference on Military Technologies	1
10th International Conference on Cloud Computing, Data Science and Engineering	1	2020 IEEE Calcutta Conference	1
10th International Conference on Intelligent Systems: Theories and Applications	1	2020 International Conference on Data Science and Its Applications	1
10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing	1	2020 International Conference on Information Management and Technology	1
10th International Conference on Security of Information and Networks	1	2020 Moratuwa Engineering Research Conference	1
11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management	1	22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	1

12th International Conference for Internet Technology and Secured Transactions	1	24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems	1
12th International Conference on Management of Digital EcoSystems	1	26th International Conference on Software, Telecommunications and Computer Networks	1
13th International Conference on Signal-Image Technology & Internet-Based Systems	1	2nd International Conference on Data, Engineering and Applications	1
13th International Workshop on Semantic and Social Media Adaptation and Personalization	1	2nd International Conference on Informatics and Computing	1
15th Iberian Conference on Information Systems and Technologies	1	2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice and 2nd International Workshop on Finance and Economics on the Semantic Web	1
16th IEEE International Symposium on Computational Intelligence and Informatics	1	2th International Conference on Digital Forensics	1
17th International Conference on Distributed Computing and Networking	1	31st International Conference on Advanced Information Networking and Applications Workshops	1
18th European Conference on Cyber Warfare and Security	1	33rd Annual Computer Security Applications Conference	1
19th International Symposium on Communications and Information Technologie	1	35th Annual ACM Symposium on Applied Computing	1
1st International Conference on Informatics and Computational Sciences	1	37th Australasian Computer Science Conference	1
1st International Informatics and Software Engineering Conference	1	3rd International Conference on Information Systems Security and Privacy	1
2014 IEEE International Conference on Systems, Man, and Cybernetics	1	48th Hawaii International Conference on System Sciences	1
2014 IEEE Joint Intelligence and Security Informatics Conference	1	49th Hawaii International Conference on System Sciences	1
2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	1	4th International Conference and Workshops on Recent Advances and Innovations in Engineering: Thriving Technologies	1
2014 International Conference on Signal Processing and Multimedia Applications	1	4th International Conference on Big Data Innovations and Applications	1

2015 European Intelligence and Security Informatics Conference	1	5th International Conference on Information Systems Security and Privacy	1
2015 IEEE International Conference on Automation Science and Engineering	1	5th International Conference on Sustainable Information Engineering and Technology	1
2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	1	6th International Conference on Reliability, Infocom Technologies and Optimization	1
2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing	1	6th International Conference on Wireless and Telematics	1
2016 IEEE Symposium Series on Computational Intelligence	1	7th Symposium on Hot Topics in the Science of Security	1
2016 International IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress	1	9th European Workshop on System Security	1
2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference	1	ACM Symposium on Document Engineering 2018	1
2017 Federated Conference on Computer Science and Information Systems	1	CEUR Workshop Proceedings	1
2017 IEEE International Conference on Big Data	1	IEEE 2nd International Conference on Information and Computer Technologies	1
2017 IEEE Symposium Series on Computational Intelligence	1	IEEE 6th International Conference on Advanced Computing	1
2017 IEEE/WIC/ACM International Conference on Web Intelligence	1	IEEE 7th Annual Computing and Communication Workshop and Conference	1
2017 Intelligent Systems Conference	1	II International Conference on Network, Communication and Computing	1
2017 International Conference on Big Data Analytics and Computational Intelligence	1	International Conference on Informatics and Analytics	1
2017 International Conference on Energy, Communication, Data Analytics and Soft Computing	1	International Joint Conference on Neural Networks	1

2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing	1	Joint 16th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 11th IEEE International Conference on Big Data Science and Engineering, and 14th IEEE International Conference On Embedded Software And Systems	1
2018 Cyber Resilience Conference	1	The International Conference on Engineering & MIS 2015	1
2018 IEEE Conference on Big Data and Analytics	1	The Web Conference 2020	1
Total 1	47	Total 2	42
		Total 1 + Total 2	**89**

Table A3. Complete list of books with chapters about text mining in public security published between 2014 and 2021.

Book	Count
Advances in Intelligent Systems and Computing	3
Advances in Computing and Data Sciences	1
Advances in Ubiquitous Networking	1
Artificial Intelligence Applications and Innovations	1
Computational Linguistics and Intelligent Text Processing	1
Data Science and Social Research	1
Databases Theory and Applications	1
Hybrid Intelligent Systems	1
Information and Communication Technology for Intelligent Systems	1
Integrated Formal Methods	1
Intelligent Computing, Communication and Devices	1
Intelligent Information Processing VIII	1
Lecture Notes in Computer Science	1
Lecture Notes in Networks and Systems	1
Mobile Networks and Management	1
Social, Cultural, and Behavioral Modeling	1
Transactions on Computational Science XXXII	1
Total	**19**

Table A4. Works related to the techniques in the 20 most frequent terms.

Technique	Works
Support Vector Machines	(Aghababaei & Makrehchi; 2018, AL-Saif & Al-Dossari, 2018; Almehmadi et al., 2017; Alothman & Rattadilok, 2017; Ariffin et al., 2018; Bhardwaj & Gupta, 2018; Chen et al., 2017; Chen et al., 2016a; Choudhary & Vidyarthi, 2015; Cichosz, 2018; Dong, Liao, et al., 2016; Dong, Xu, et al., 2016; Elkhawas & Abdelbaki, 2018; Hajek & Henriques, 2017; Husari et al., 2017; Li, Xu, et al., 2014; Maktabar et al., 2018; Mine et al., 2019; Mohasseb et al., 2019; Noviantho et al., 2017; Parapar et al., 2014; Petrovskiy & Chikunov, 2019; Saini & Bansal, 2019; Samtani et al., 2017; Samtani et al., 2018; Sharmin & Zaman, 2017; Sonowal & Kuppusamy, 2018; ; Spitters et al., 2015; Sundarkumar et al., 2015; Thao et al., 2017; Venckauskas et al., 2017; Vidyarthi et al., 2017; Wajid & Samet, 2016; Zareapoor & Seeja, 2015; Zhao et al., 2016; Alatrista-Salas et al., 2020; Andleeb et al., 2019; Balim & Gunal, 2019; Battaglia et al., 2020; Bozyiğit et al., 2021; Castillo-Zúñiga et al., 2020; Gravanis et al., 2019; Monish & Pandey, 2020; Rabuzin & Modrušan, 2019; Samtani et al., 2020; Siering et al., 2021; Wei et al., 2020; Rehman et al., 2020; Miranda et al., 2020; Ishara Amali & Jayalal, 2020; Pires & Georgieva, 2020; Sahu et al., 2019)
Term Frequency - Inverse Document Frequency	(Aghababaei & Makrehchi, 2018; Agrawal & Kaushal, 2016; AL-Saif & Al-Dossari, 2018; Alothman & Rattadilok, 2017; Anwar & Abulaish, 2014a; Ariffin et al., 2018; Chen et al., 2017; Chen et al., 2016a; Chen et al., 2016b; Chung et al., 2018; Dong et al., 2018; Dong, Liao, et al., 2016; Dong, Xu, et al., 2016; Fa et al., 2017; Husari et al., 2017; Kuang et al., 2017; Lekea & Karampelas, 2017; Mansour, 2018; Martin et al., 2016; Netsuwan & Kesorn, 2017; Nikolić et al., 2015; Noviantho et al., 2017; Parapar et al., 2014; Percy et al., 2018; Qazi & Wong, 2019; Sharmin & Zaman, 2017; Suarez; Tangil et al., 2014; Sundarkumar et al., 2015; Thao et al., 2017; Xianghui et al., 2015; Yang et al., 2016; Zahra et al., 2018; Zhao et al., 2016; Alatrista-Salas et al., 2020; Balim & Gunal, 2019; Battaglia et al., 2020; Bozyiğit et al., 2021; Castillo-Zúñiga et al., 2020; de Boer et al., 2019; Gomes & Ladeira, 2020; Kabwe & Phiri, 2020; Lyu et al., 2020; Rabuzin & Modrušan, 2019; Siering et al., 2021; Rehman et al., 2020; Kakavand et al., 2019; Lal et al., 2020; Pires & Georgieva, 2020)
Naïve Bayes	(Al-Khalisy and Jehlol, 2018; AL-Saif & Al-Dossari, 2018; Alothman & Rattadilok, 2017; Bhardwaj & Gupta, 2018; Chen et al., 2016a; Cichosz, 2018; Fa et al., 2017; Hajek & Henriques, 2017; Maktabar et al., 2018; Martin et al., 2016; Mohasseb et al., 2019; Netsuwan & Kesorn, 2017; Noviantho et al., 2017; Overbeck, 2015; Samtani et al., 2018; Sharef et al., 2015; Sharmin & Zaman, 2017; Thao et al., 2017; Tran & Tran, 2018; Xia et al., 2019; Zahra et al., 2018; Zareapoor & Seeja, 2015; Adikara et al., 2020; Andleeb et al., 2019; Bozyiğit et al., 2021; Castillo-Zúñiga et al., 2020; Malim et al., 2019; Gravanis et al., 2019; Mukherjee & Sarkar, 2020; Rabuzin & Modrušan, 2019; Wei et al., 2020; Rehman et al., 2020; Sudha & Rupa, 2019; Ishara Amali & Jayalal, 2020; Lal et al., 2020; Palad et al., 2019; Sahu et al., 2019; Saldana et al., 2020)
Random Forests	(Alothman & Rattadilok, 2017; Cichosz, 2018; Fa et al., 2017; Fontanarava et al., 2017; Hadad et al., 2017; Hadad et al., 2018; Hajek & Henriques, 2017; Kim et al., 2018; Marivate & Moiloa, 2016; Martin et al., 2016; Petrovskiy & Chikunov, 2019; Saini & Bansal, 2019; Samtani et al., 2018; Sonowal & Kuppusamy, 2018; Sundarkumar et al., 2015; Zareapoor & Seeja, 2015; Angenent et al., 2020; Balim & Gunal, 2019; Battaglia et al., 2020; Bozyiğit et al., 2021; Castillo-Zúñiga et al., 2020; Rehman et al., 2020; Lal et al., 2020; Palad et al., 2020; Pires & Georgieva, 2020)

Decision Trees	(AL-Saif & Al-Dossari, 2018; Alothman & Rattadilok, 2017; Bhardwaj & Gupta, 2018; Chen et al., 2016a; Hadad et al., 2017; Hadad et al., 2018; Hajek & Henriques, 2017; Li, Xu, et al., 2014; Martin et al., 2016; Netsuwan & Kesorn, 2017; Petrovskiy & Chikunov, 2019; Saini & Bansal, 2019; Sharef et al., 2015; Sonowal & Kuppusamy, 2018; ; Sundarkumar et al., 2015; Thao et al., 2017; Battaglia et al., 2020; Calderon et al., 2020; Castillo-Zúñiga et al., 2020; Malim et al., 2019; Gravanis et al., 2019; Monish & Pandey, 2020; Palad et al., 2020; Palad et al., 2019; Pires & Georgieva, 2020)
Logistic Regression	(Aboluwarin et al., 2016; Chen et al., 2016a; Chen et al., 2016b; Cichosz, 2018; Fontanarava et al., 2017; Hadad et al., 2017; Hadad et al., 2018; Hajek & Henriques, 2017; Maktabar et al., 2018; Marivate & Moiloa, 2016; Parapar et al., 2014; Petrovskiy & Chikunov, 2019; Ristea et al., 2018; Saha et al., 2016; Samtani et al., 2018; Balim & Gunal, 2019; Bozyiğit et al., 2021; Monish & Pandey, 2020; Rabuzin & Modrušan, 2019; Wei et al., 2020; Pires & Georgieva, 2020; Sahu et al., 2019)
Named Entity Recognition	(Anwar & Abulaish, 2014a; Ariffin et al., 2018; Badii et al., 2014; Behmer et al., 2019; Bisgin et al., 2019; Chen et al., 2017; Das & Das, 2017a; Das & Das, 2019; Iftikhar et al., 2019; Seidler et al., 2014; Wajid & Samet, 2016; Yang et al., 2016; Yang et al., 2014; Zaeem et al., 2017; Calderon et al., 2020; de Boer et al., 2019; Tajuddin et al., 2019; Palad et al., 2020; Palad et al., 2019)
Latent Dirichlet Allocation	(Aghababaei & Makrehchi, 2018; Basilio et al., 2019; Bisgin et al., 2019; Dong, Liao, et al., 2016; Dong, Xu, et al., 2016; Lee et al., 2016; Li et al., 2016; Marivate & Moiloa, 2016; Noel & Peterson, 2014; Overbeck, 2015; Ristea et al., 2018; Samtani et al., 2017; Sundarkumar et al., 2015; Yang et al., 2016; Alagheband et al., 2020; Basilio et al., 2020; Al-Ramahi et al., 2020; Birks et al., 2020)
k-Nearest Neighbors	(AL-Saif & Al-Dossari, 2018; Alothman & Rattadilok, 2017; Andriansyah et al., 2018; Barbon Jr. et al., 2017; Chandra et al., 2017; Chen et al., 2016a; Martin et al., 2016; Subhan et al., 2017; Battaglia et al., 2020; Bozyiğit et al., 2021; Gravanis et al., 2019; Monish & Pandey, 2020; Ishara Amali & Jayalal, 2020; Kaur et al., 2019; Pires & Georgieva, 2020)
Manual Annotation	(Alakrot et al., 2018; Andriansyah et al., 2018; Dong et al., 2018; Iftikhar et al., 2019; Johnston & Weiss, 2017; Karystianis et al., 2018; Karystianis et al., 2019; Liang et al., 2016; Liang & Biros, 2016; Petrovskiy & Chikunov, 2019; Saha et al., 2016; Saini & Bansal, 2019; Samtani et al., 2017; Subhan et al., 2017; Ishara Amali & Jayalal, 2020)
Cosine Similarity	(Chen et al., 2017; Chen et al., 2016a; Das & Das, 2017c; Kuang et al., 2017; Lekea & Karampelas, 2017; Nedeljkovic et al., 2019; Percy et al., 2018; Qazi & Wong, 2019; Spitters et al., 2015; Suarez-Tangil et al., 2014; Xianghui et al., 2015; Zhao et al., 2016; Gomes & Ladeira, 2020; Kabwe & Phiri, 2020)
Dictionaries	(Karystianis et al., 2018; Karystianis et al., 2019; Liang et al., 2016; Liang & Biros, 2016; Mansour, 2018; Miah et al., 2015; Nwafor et al., 2016; Percy et al., 2018; Po & Rollo, 2018; Roopa & Induja, 2019; Wajid & Samet, 2016)
Term Frequency	(Cichosz, 2018; Fa et al., 2017; Gil et al., 2018; Hao & Dai, 2016; Mohasseb et al., 2019; Niekerk et al., 2019; Öztürk & Ayvaz, 2018; Samtani et al., 2017; Sonowal & Kuppusamy, 2018; Zainal et al., 2018; Castillo-Zúñiga et al., 2020)
Neural Networks	(Fontanarava et al., 2017; Johnston & Weiss, 2017; Martinelli et al., 2017; Netsuwan & Kesorn, 2017; Samtani et al., 2018; Thao et al., 2017; Alatrista-Salas et al., 2020; Castillo-Zúñiga et al., 2020; Monish & Pandey, 2020)

Table A5. Works related to the technologies in the 20 most frequent terms.

Technology	Works
Python	(Aboluwarin et al., 2016; Cardoza & Wagh, 2017; Chen et al., 2016a; Chung et al., 2018; Das & Das, 2017a; Das & Das, 2017b; Das & Das, 2017c; Das & Das, 2019; Dong et al., 2018; Elkhawas & Abdelbaki, 2018; Hernandez-Castro & Roberts, 2015; Husari et al., 2018; Lee et al., 2016; Lekea & Karampelas, 2017; Maktabar et al., 2018; Marivate & Moiloa, 2016; Martin et al., 2016; Martinelli et al., 2017; Nedeljkovic et al., 2019; Nguyen et al., 2017; Samtani et al., 2018; Savaş & Topaloğlu, 2019; Sonowal & Kuppusamy, 2018; ; Thao et al., 2017; Vidyarthi et al., 2017; Yang et al., 2016; Alagheband et al., 2020; Al-Nabki et al., 2020; Andleeb et al., 2019; Angenent et al., 2020; Birks et al., 2020; Bozyiğit et al., 2021; de Boer et al., 2019; Gomes & Ladeira, 2020; Hou et al., 2020; Malim et al., 2019; Rabuzin & Modrušan, 2019; Samtani et al., 2020; Wei et al., 2020; Miranda et al., 2020; Ishara Amali & Jayalal, 2020; Pires & Georgieva, 2020; Saldana et al., 2020)
R language	(Basilio et al., 2019; Bisgin et al., 2019; Cataldo et al., 2017; Chandra et al., 2017; Chang & Wang, 2015; Cichosz, 2018; Correa et al., 2018; Elkhawas & Abdelbaki, 2018; Gil et al., 2018; Hao & Dai, 2016; Hultgren et al., 2018; Mansour, 2018; Martin et al., 2016; Öztürk & Ayvaz, 2018; Pina-Sánchez et al., 2019; Dastjerdi et al., 2019; Ristea et al., 2018; Roopa & Induja, 2019; Saini & Bansal, 2019; Silomon & Roeling, 2018; Basilio et al., 2020; Castillo-Zúñiga et al., 2020)
WEKA	(Almehmadi et al., 2017; Alothman & Rattadilok, 2017; Choudhary & Vidyarthi, 2015; Das & Das, 2017b; Das & Das, 2017c; Ding et al., 2015; Hadad et al., 2017; Hadad et al., 2018; Kim et al., 2018; Li, Xu, et al., 2014; Samtani et al., 2018; Sharef et al., 2015; Sharmin & Zaman, 2017; Ventirozos et al., 2018; Vidyarthi et al., 2017; Calderon et al., 2020; Lal et al., 2020; Palad et al., 2020; Palad et al., 2019)
Scikit-Learn	(Aboluwarin et al., 2016; Chen et al., 2016a; Chung et al., 2018; Dong et al., 2018; Husari et al., 2018; Lekea &; Karampelas, 2017; Martin et al., 2016; Thao et al., 2017; Angenent et al., 2020; Bozyiğit et al., 2021; Rabuzin & Modrušan, 2019; Samtani et al., 2020; Miranda et al., 2020; Ishara Amali & Jayalal, 2020)
Natural Language Toolkit	(Aboluwarin et al., 2016; Cardoza & Wagh, 2017; Chung et al., 2018; Das & Das, 2019; Dong et al., 2018; Lee et al., 2016; Marivate & Moiloa, 2016; Sonowal & Kuppusamy, 2018; Yang et al., 2016; Andleeb et al., 2019; de Boer et al., 2019; Samtani et al., 2020; Miranda et al., 2020)
RapidMiner	(AL-Saif & Al-Dossari, 2018; Bhardwaj & Gupta, 2018; Margono et al., 2014; Noviantho et al., 2017; Samtani et al., 2017; Sundarkumar et al., 2015; Tayal & Ravi, 2016; Zainal et al., 2018; Saldana et al., 2020)

Table A6. The most frequent techniques and technologies terms in the Cybersecurity area.

Extracted techniques/technologies terms	**Count**
python	17
support vector machines	16
decision trees	13
term frequency-inverse document frequency	20
random forests	12
naïve bayes	11

weka	8
k-nearest neighbors, logistic regression, scikit-learn, term frequency	7
natural language toolkit, r language	6
adaboost, named entity recognition, word clouds, support vector machines	4

Table A7. The most frequent techniques and technologies terms in the General crime detection/prediction area.

Extracted techniques/technologies terms	**Count**
python	10
naïve bayes, term frequency-inverse document frequency	8
support vector machines	6
weka, cosine similarity	5
decision trees, latent dirichlet allocation, named entity recognition	4
euclidean similarity, jaccard similarity, dictionaries	3
cluster analysis, georeferencing, logistic regression, natural language toolkit, neural networks, random forests, rapidminer	2

Table A8. The most frequent techniques and technologies terms in the Fraud detection area.

Extracted techniques/technologies terms	**Count**
support vector machines	11
logistic regression	7
term frequency-inverse document frequency	6
naïve bayes	5
decision trees	4
random forests, python, named entity recognition	3
bagging, georeferencing, latent dirichlet allocation, loss calculation, matlab, neural networks, risk calculation, scikit-learn, cosine similarity, principal component analysis	2

Table A9. The most frequent techniques and technologies terms in the Terrorism detection area.

Extracted techniques/technologies terms	**Count**
r language, naïve bayes, term frequency-inverse document frequency	5
support vector machines	4
network graphs, random forests, manual annotation	3
decision trees, latent dirichlet allocation, neural networks, scikit-learn, term frequency	2

Table A10. The most frequent techniques and technologies terms in the Cyberbullying detection area.

Extracted techniques/technologies terms	**Count**
naïve bayes, support vector machines	5

k-nearest neighbors, term frequency-inverse document frequency, python	3
libsvm, manual annotation, rapidminer, scikit-learn	2

Table A11. The most frequent techniques and technologies terms in the Digital/Cyber forensics area.

Extracted techniques/technologies terms	**Count**
latent dirichlet allocation	3
natural language toolkit, python, support vector machines, named entity recognition	2

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.