An Automated Corpus Annotation Experiment in Brazilian Portuguese for Sentiment Analysis in Public Security

Victor Diogho Heuer de Carvalho[1,2(✉)], Thyago Celso Cavalcante Nepomuceno[3], and Ana Paula Cabral Seixas Costa[2]

[1] Universidade Federal de Alagoas, Delmiro Gouveia, AL, Brazil
[2] Universidade Federal de Pernambuco, Recife, PE, Brazil
victor.hcarvalho@ufpe.br, apcabral@cdsid.org.br
[3] Universidade Federal de Pernambuco, Caruaru, PE, Brazil
thyago.nepomuceno@ufpe.br

Abstract. This paper aims to present an experiment developed in order to produce a corpus with automated annotation, using pre-existing annotated corpus and machine learning classification methods. A search for pre-existing annotated corpora in Brazilian Portuguese was applied, founding six corpora of which one has been selected as the training dataset. A set of tweets was collected in a specific area of Recife (Pernambuco-Brazil) using some keywords related to kinds of crimes and reinforcing some places in that area. Preprocessing tasks were applied over the pre-existing corpus and the tweets' set collected. Latent Dirichlet Allocation was applied for topic modeling followed by Multinomial Naïve Bayes, Linear Support Vector Machines, and Logistic Regression for the sentiment polarity classification. The results of the cross-validation of the experiment indicated Linear Support Vector Machines as the most accurate classification method among the three considering the specific training set used, and by this method, the new annotated corpus about the selected topic related to public security was created.

Keywords: Corpus annotation · Sentiment analysis · Public security · Brazilian Portuguese · Machine learning classification

1 Introduction

The social web has added a new level of challenge for organizations that want to collect data from this massive source to apply in their decision-making [1]. Big data and social media analytics have become necessary concepts for companies to remain competitive while meeting the expectations of their customers/consumers [2]. From these concepts arises the need to analyze the massive and heterogeneous volumes of data from the social web [3] in order to provide more accurate information based on public opinion decision-makers [4].

Daily social network users produce textual records that can be useful for organizations if properly handled, and text mining is essential to ensure these records can generate

J. M. Moreno-Jiménez et al. (Eds.): ICDSST 2020, LNBIP 384, pp. 99–111, 2020.
https://doi.org/10.1007/978-3-030-46224-6_8

useful information [5, 6]. Web mining is a text mining approach that aims to explore and extract data from various sources across the internet, enabling pattern recognition and the extraction of useful information [7]. As an intermediate outcome, sentiment analysis, which in turn is a specific type of natural language processing problem, can be used to classify the polarity of textually expressed opinions [4, 8].

Public services management agencies are potential users of the tools and approaches mentioned above, highlighting applications in health, environment, and security [9]. Public security agencies, for instance, can use social web mining to extract people's impressions about security policies and actions from social networks aggregating these information to reports on crimes and policing in specific periods and geographical regions, assisting decisions on measures to improve these policies and actions [10] or crossing the results of the analytical process with internal data to support decisions on improving surveillance by predicting events against security [11].

Assessing people's sentiments about public security is a process of interest to government public security management agencies. General and specific misunderstandings have been recurring topics on the authorities' agenda, and many recent studies are devoted to discussing the empirical context of crime in the environment in which such data collection is done [12–14].

Sentiment analysis requires a series of preparations so that the opinions expressed by people can be classified appropriately and reliably to the sentiment they really wanted to express in their textual records [15]. The creation and use of an adequately annotated corpus are fundamental for the classification of sentiments polarities as the main result of the analytical process [16].

The annotation process can be done manually, using individuals with enough expertise to analyze the texts and apply a polarity label [17, 18], or using automated procedures based on machine learning techniques which depends on a pre-existing annotated corpus and computational capacity to process the new corpus in a real-time [19, 20].

Topic modeling is another important task to ensure obtaining a specific domain corpus. It allows the most recurring words in the text to be identified and clustered so that each cluster defines a topic that can also be annotated in a corpus, allowing the extraction of specific textual records for a topic of interest [21].

This paper aims to present a corpus annotation experiment dedicated specifically to a process of sentiment analysis about Public Security in the city of Recife (Brazil). To this end, a set of tweets was collected in a specific area of the city, using some key terms related to crimes and reinforcing some places in that area. Topic modeling was performed to identify the tweets most adjusted to the public security theme and, afterward, an automated procedure was applied using a pre-existing corpus for the annotation of the new (specialized) corpus. Obtaining a specialized annotated corpus is a necessary process to work on a specific domain, as occurs on the broader context in which this paper is inserted.

The sequence is divided as follows: Sect. 2 provides a background on the corpus annotation process; Sect. 3 presents the annotation experiment process applied; Sect. 4 presents the experiment results; lastly, Sect. 5 presents the concluding remarks containing some practical implications and indications of future work.

2 Background

The following subsections introduce some main concepts about the automated corpus annotation process, topic modeling, and sentiments polarity labeling.

2.1 Automated Corpus Annotation

Automated corpus annotation presupposes the use of a pre-existing annotated and general-purpose corpus as a training dataset for machine learning algorithms to perform polarity labeling over a new domain-specific corpus [22].

It is important to emphasize that neither manual and automated processes are infallible and both have advantages and disadvantages [23]: (i) manual annotation can be executed with little corpus preparation, and ensure the best accuracy in the results since it is a process-oriented to human perceptions; in contrast, the process is slow and limited to few results; (ii) automated annotation can work with broader corpus and the process is much faster compared to manual, but it involves the programming of a labeling or tagging process using artificial intelligence techniques that should be first trained, so, the accuracy of the results will depend on the quality of the pre-existing training set.

For automated annotation, sometimes, several pre-existing corpora can be used to ensure more accurate classifications of the sentiments [20]. Besides, some review and re-annotation processes may be necessary to ensure the quality of the new corpus [19]. Preprocessing techniques also are required for the preparation of the corpus as well as the application of cross-validation using metrics like accuracy, precision, recall, and F1-score [18, 20, 24, 25].

2.2 Topic Modeling

The topic modeling allows the classification/labeling of texts according to the topics found [26], ensuring the identification of discussion subjects in the texts [27], supporting the creation of a new corpus.

Latent Dirichlet Allocation (LDA) figures as a recurring topic modeling technique in recent literature (see, for instance, [27–30]). It belongs to the class of hierarchical Bayesian models describing documents as a mixture of topics [29] and still can be classified as an unsupervised machine learning generative technique [31]. LDA was developed to fix issues related to Latent Semantic Analysis (LSA) and its probabilistic version (PLSA) [32], two other topic modeling techniques based on the use of a general matrix of texts and terms to be decomposed in other two matrixes with the relations "document to topic" and "topic to term" [31].

2.3 Sentiment Polarity Labeling

Sentiment polarity labeling is a process that can be performed by machine learning techniques, lexicon-based methods, or hybrid approaches [33]. The machine learning techniques applied to sentiment analysis are designated to train text classifiers, according to pre-existing datasets with polarity annotation [25].

The most recurrent machine learning techniques for this purpose on literature are the classic trio Naïve Bayes, Support Vector Machines (SVM) and Maximum Entropy/Logistic regression followed by artificial neural networks and deep learning [34], but other supervised methods such as Random Forests, Nearest Neighbors, as well as unsupervised methods as K-Means [33], for instance, can be applied.

Lexicon based methods involve the calculation of the text semantic orientation to determine the polarity [16]. The approach involved on this process is oriented to counting and weighting the sentiment-related words, and it can be performed by three ways [25]: (i) a manual process, as mentioned before, depends on human beings and is time-consuming; (ii) a dictionary-based process explores lexicographical resources, as WordNet, for instance; (iii) a corpus-based process uses sets of words with well-defined sentiment polarity exploring syntactic relations to identify new sentiment words.

Lastly, hybrid approaches may combine both machine learning/statistics methods and lexicon-based approaches [16], helping to improve the accuracy of sentiments classification [35, 36]. The next section will present the steps of the experiment applied to perform corpus annotation using a pre-existing annotated corpus in Brazilian Portuguese, and tweets collected from a specific area in a large Brazilian city.

3 Experiment Description

The corpus annotation experiment adopted in this work was based on automated tasks. It involved as the first task the search for a pre-existing annotated general-purpose corpus in Brazilian Portuguese and the collection of a set of tweets in a specific area from a Brazilian city, using some initial filtering based on keywords and setting the collector only to get tweets in the target language.

The next task was to apply topic modeling to identify which topics were, in fact, relevant for obtaining the corpus dedicated to public security sentiment analysis. Lastly, using the pre-existing corpus and the tweets extracted based on topic modeling, sentiment analysis was applied to obtain the final specific annotated corpus in Brazilian Portuguese. Figure 1 presents the workflow of this process.

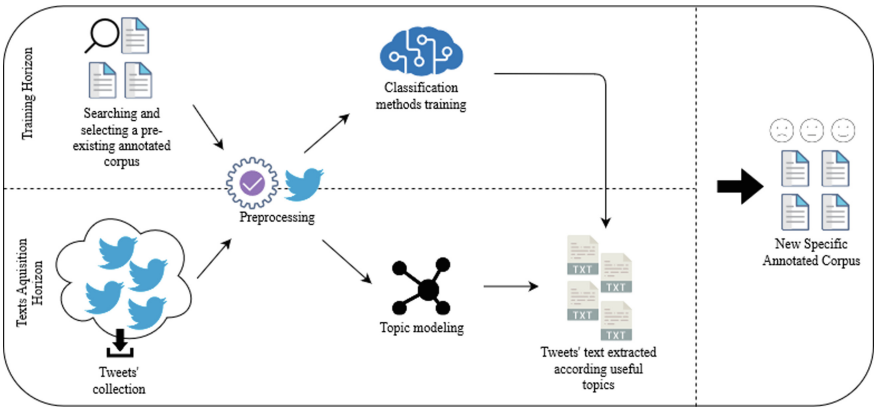

Fig. 1. Annotation experiment process workflow.

Each element in this workflow will be described in the following subsections, demonstrating what technologies and methods were used.

3.1 Searching and Selecting a Pre-existing Annotated Corpus

Here, pre-existing annotated corpora obtained through well-structured processes were sought to select a textual training data set for machine learning algorithms applied in the classification of the sentiments' polarities. This search found the corpora related to the following works: [17, 37–41]. One of the corpora [38] was discarded for not presenting sentiments polarity labeling. Another corpus [39] was based on a bookshelf only applied positive and negative labels, being eliminated as it was desired to consider a neutral label too. Among the remaining corpus, the one related to the UniLex method [40] was chosen, since it fits well with the classification proposal.

The selected corpus consisted of fifteen XLSX files that were read through a Python script using the Glob library to read multiple files and Pandas to load the data from these files into a data frame and convert this structure to a final file in CSV format. The texts on the final file were preprocessed by a function that uses some methods from Natural Language Toolkit (NLTK) [42] and "re" (regular expressions) libraries for data cleaning and stop words removal, using Portuguese stop words downloaded via NLTK. The preprocessing tasks ensure noises elimination in the training of the classification methods. More details about the selected corpus are presented in the results section.

3.2 Tweets' Collection and Preprocessing

Tweet's collection was based on the procedure to collect users' posts on Twitter's continuous stream, storing them in a JSON file [43]. A stream collector script was implemented using Python language applying the following libraries and their methods: Tweepy for accessing the Twitter API, "time" providing function for dealing with time elements, and "os" for miscellaneous interface with the operating system. The JSON content was charged in Pandas' data frames to be manipulated by other scripts.

Preparing the tweets' data for the topic modeling and polarity labeling tasks is an element of significant interest in the whole analytical process since its focus is to give a suitable structure so that the texts can be effectively analyzed. For this purpose, duplicate tweets were dropped, and the same preprocessing function applied to the selected corpus was used to preprocess the tweets' texts also to eliminate noise for the subsequent parts of the analysis.

3.3 Topic Modeling

LDA was the method selected for topic modeling, following the tendency presented in the background section. Reading the works in the previous section related to LDA applications is recommended to ensure understanding of how the method works. A new Python script was implemented to apply topic modeling, this time using Scikit-Learn library calling functions for feature selection (TF-IDF, term frequency-inverse document frequency), LDA functions, and Grid Search functions to select the best topic model according to LDA results and the feature selection applied.

Another scientific programming and graph plotting supporting libraries like Numpy, Matplotlib, and Seaborn also were used. Plotting libraries allowed the visualization of the most recurrent textual features selected (bigrams), and of the topics segmentation as texts' clusters. The k-Means technique, from Scikit-Learn, was used here for the sole purpose of segmenting these topics into the related texts according to the labeling done by the LDA, generating the graph with the corresponding visualization.

With topic modeling, the tweets could be labeled based on the best topic model found, and the final selection of tweets searched for texts related to public security issues occurred, defining the subset of texts for polarity labeling.

3.4 Polarity Labeling Over the Selected Tweets

The polarity labeling is the final part of the procedure, dedicated to classifying the sentiments' polarities using some supervised machine learning classification techniques. The steps presented in Subsects. 3.1, 3.2, and 3.3 provided (i) a pre-existing annotated corpus, (ii) a broader set of unclassified tweets, and (iii) a subset of unclassified tweets extracted according to a public security-related topic identified using LDA.

The pre-existing annotated corpus contains the sentiments' polarities manually annotated using three labels: negative (−1), neutral (0), and positive (1). It was used to train the machine learning techniques, applying cross-validation, and allowing to evaluate the sentiment classification according to specific metrics [25] supporting discovering which is the most suitable technique for this task among those used. It is important to emphasize that only the tweets in the subset extracted using LDA were used for applying the automated polarity labeling/annotation, using the classification techniques.

The last Python script was implemented to process polarity labeling, using the three classification techniques from Scikit-Learn: Multinomial Naïve Bayes, Linear Support Vector Machines (Linear SVM), and Logistic Regression. The results of the described process will be presented following.

4 Experiment Results

4.1 Pre-existing Annotated Corpus Characteristics

The pre-existing corpus, derived from the UniLex method [40], is composed of 12668 tweets extracted in Brazilian Portuguese, related to issues about national politics. It is openly available to be used in sentiments analysis processes. The corpus structure is composed of 3 columns, the first containing the index o the tweets, the second the tweets, and the third the labels classifying the tweets (negative, neutral, and positive). The texts are distributed in 4197 negatives, 4753 neutrals, and 3715 positive registers.

No duplicate entries were found, however, three invalid entries were found and needed to be deleted, resulting in an amount of 12665 tweets with the same amounts for each polarity label. Table 1 exemplifies the first three registers of the corpus after preprocessing with tweets' texts in Brazilian Portuguese (index column was added in the initial corpus to data frame conversion).

Data preprocessing eliminated special characters, the "RT" (retweet) indication, entity tags with "@", emojis and stop words from Brazilian Portuguese, reducing noises with unnecessary textual elements during the training of the classification methods.

Table 1. Example of the corpus structure.

Index	Tweets	Labels
0	#caonossodecadadia #novo vanessa mandotti dir pgm começo	Neutral (0)
1	bola frente amanhã outro dia, outra cena, outra chance, viver ser feliz irmão amém #amanha #novo #viver	Positive (1)
2	cara mal? acho apenas corte diferente barba #visu #novo #gostou	Positive (1)

4.2 Extracted Tweets Characteristics

The stream tweets collector script was parameterized to perform the searches in a specific quadrant of the city of Recife, on the following coordinates: −34.9090 for west latitude, −8.0716 for south longitude, −34.8753 for east latitude and −8.0522 for north longitude. In addition to these geographical delimiters, the specific language of the collection was defined to Portuguese, and some keywords were introduced for filtering the initial collection: *["tráfico de drogas", "estupro", "homicídio", "assassinato", "furto", "roubo", "assalto", "Soledade", "Derby", "Boa Vista", "Madalena", "Ilha do Retiro", "Paissandu", "Ilha do Leite", "Joana Bezerra", "Santo Antônio", "Santo Amaro", "Coque", "Borel", "Ilha de Deus", "Antigo", "Pina"]*. These keywords are related to some common crimes' designations in Portuguese and some places, to reinforce the tweets' searches related to the defined Recife's quadrant.

The number of tweets collected was 17218, but after duplicities exclusion, the number was reduced to 10283. These tweets were stored in a JSON file, and again, the same preprocessing tasks used before were applied.

4.3 Topic Modeling Results

After preprocessing the tweets collected on the defined Recife's quadrant, the topic modeling process was performed. TF-IDF using bigrams was applied instead of simple counting, to create the bag-of-words and discover the frequency of the bigrams among the tweets' texts. Figure 2 contains a bar plot with the ten most common bigrams found on the text set.

A Grid Search was applied to find the best LDA model according to the number of topics in an analysis using 5, 10, 15, 20, 25, and 30 topics, and 0.5, 0.7, and 0.9 learning decays. The results indicated the best log-likelihood score (−144000.0555) and model perplexity (408562.2158) for a model with five topics and a learning decay of 0.7. As higher the log-likelihood and as lower the model perplexity, the best the model [44]. The model was formulated to contain the ten most relevant bigrams extracted from the tweet texts for each topic.

The topics in the texts' set follow the distribution presented in Table 2, where it can be noted that topic 0 is the most frequent. The occurrence numbers in this table were determined using the function *values_count* from Pandas (Python's library), according to the topics extracted through LDA.

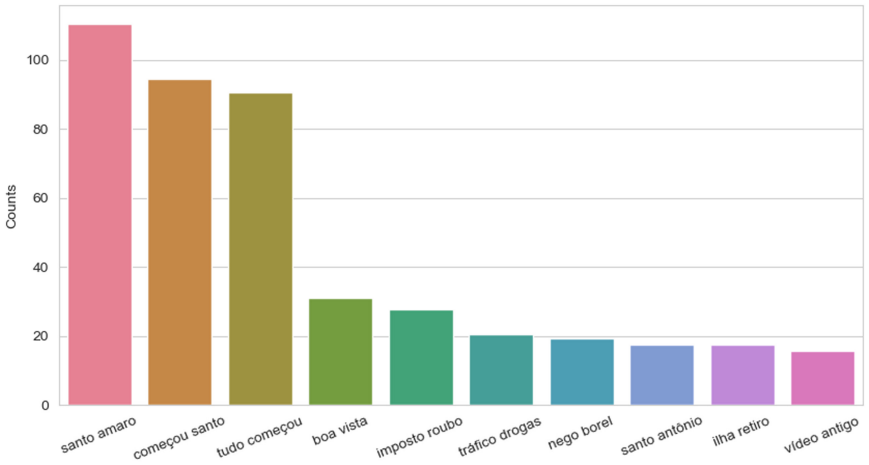

Fig. 2. The ten most common bigrams on tweets' texts.

Table 2. The occurrence numbers of each topic on the model.

Topic	Occurrences
0	2278
3	2151
4	1997
1	1958
2	1899
Total	10283

Figure 3 presents the segmentation of the topics (clusters) on the tweets' set, according to the components' weights, using the K-Means technique.

Finally, only tweets related to "Topic 0" were extracted in a new data frame and converted to a CSV file, to proceed with the last part of the experiment. This topic seems to be related to some crimes (robbery, murder, and drug trafficking) that took place in the vicinity of the central area of Recife. The top 10 bigrams in this topic are: *[“mulheres manifestações”, “vou fazer”, “violência contra”, “tentativa homicídio”, “ilha retiro”, “gente caso”, “boa vista”, “tráfico drogas”, “nego borel”, “amor antigo”].*

4.4 Polarity Labeling and Public Security Related Corpus Obtention

The last part of the experiment started with using the preprocessing results from the pre-existing corpus with sentiment labeling to train the three classification methods: Multinomial Naïve Bayes, Linear SVM, and Logistic Regression. Again, a bag-of-words was created using TF-IDF, but this time with unigrams.

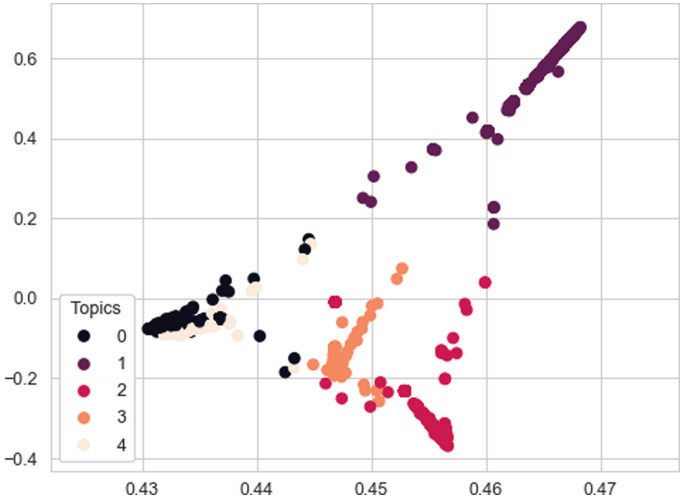

Fig. 3. Topics (clusters) segmentation.

Table 3 presented the summary of the cross-validation of the methods, using accuracy for each method in general and precision, recall, and F1-score metrics for each polarity in each method.

Table 3. Methods metrics summary related to the training corpus.

Methods	Multinomial Naïve Bayes			Linear SVM			Logistic regression		
Metric	Polarity								
	−1	0	1	−1	0	1	−1	0	1
Accuracy	0.5507			0.5533			0.5492		
Precision	0.53	0.60	0.51	0.58	0.57	0.48	0.58	0.56	0.48
Recall	0.77	0.56	0.28	0.64	0.60	0.40	0.63	0.61	0.37
F1-score	0.63	0.58	0.36	0.61	0.58	0.44	0.60	0.59	0.42

This table supported the final method selection for the corpus related to a Public Security topic annotation. Linear SVM presented the best accuracy on the cross-validation, with 0.5533 against 0.5507 from Multinomial Naïve Bayes (the second-best) and 0.5492 from Logistic Regression. Thus, for the experiment, the Linear SVM was selected as the final training method for the new corpus.

The new corpus was saved in a CSV file, and tests applying the cross-validation over it as a new training set presented an accuracy of 0.7405 for Logistic Regression, 0.6527 for Multinomial Naïve Bayes and 0.7379 for Linear SVM.

5 Concluding Remarks

This paper presented some essential concepts about automated corpus annotation for sentiment analysis and applied an experiment using some of these concepts. Sentiment analysis represents a central element in a research that is being developed with application in public security in the city of Recife (Brazil). As the main result, a corpus was obtained from the applied experiment for use as a training dataset to automate the classification of tweets' polarities about public security events, situations or occurrences with Twitter users in a specific region of the city.

The work developed has its novelty in the creation of a specific corpus for applications of sentiment analysis in public security in Portuguese language, also dedicated to a specific urban region in Brazil. The idea, therefore, is to explore the distribution of sentiment about public security in the Metropolitan Region of Recife from textual records of tweets published by people from that city, writing down a portion of these records and using it as a training set to categorize the polarities of the sentiments expressed in new texts also from this region. This does not preclude that the described procedure is used in other geographic regions and other fields such as public education, transportation, and health.

Future studies should be developed using computational power higher than that used in the reported experiment, and depending on the amount of data, prepared to deal with Big Data strategies. Other methods can be tested in both topic modeling, sentiment polarity classification, and further searches about pre-existing annotated corpora in the Portuguese language will be done. A more refined assessment of sentiment polarity classification methods will also be made, based on all metrics used, to assist in the ranking and selection of the best method.

For those interested in applying the process presented in the experiment in other languages, it is recommended to change the search language in the searching tool. Changes in the geographic region where the collection of tweets will be applied will also imply language changes, as well as the determination of key-terms for the search in the target language. For instance, in the experiment reported in this work, these parameters were determined in a Twitter stream collector script written in Python and can be easily changed to do searches in other languages or regions.

For training sets in other languages, those interested can search for corpora in the target language. Another option is the use of some automated translation process, taking NLTK, which contains translation functions (nltk.translation), as an example.

Other treatments regarding irony/sarcasm and vagueness should also be applied to improve the process of obtaining new corpora for use in research. It is also intended to use geo-referenced information from social network records, when available, for the classification of urban areas regarding the sentiments of their inhabitants about public security events and measures. Instead of using only metrics like precision, accuracy, recall, and F1-score, the Receiver Operating Characteristic (ROC) curve can be applied too, allowing the visualization of the general performances of each method, assisting in the selection of the most adjusted.

The results of the analyses carried out in this process are addressed to support activities from the Secretariat of Social Defense of Pernambuco since the research being

developed is part of an agreement between the state government and the Federal University of Pernambuco, where the involved researchers are located. The process will also be incorporated into a sentiment visualization dashboard, allowing governmental managers to visualize public sentiment distribution over time and geographic space. This dashboard was planned as a module of a decision support system that is also being developed as an outcome for the referred agreement.

The sentiment visualization dashboard should help managers to combine time and geographic location to analyze the evolution of people's opinions registered in social networks, extracted through sentiment mining and analysis, and combining it, for instance, with other internal and external information about policing and crime occurrences to discover areas that demand more considerable attention for security actions.

Acknowledgment. This paper was funded in part by the Coordination for the Improvement of Higher Education Personnel (Brazil) – Finance Code 001, and by the National Council for Scientific and Technological Development (Brazil).

References

1. He, W., Wang, F.K., Akula, V.: Managing extracted knowledge from big social media data for business decision making. J. Knowl. Manage **21**, 275–294 (2017). https://doi.org/10.1108/JKM-07-2015-0296
2. Vatrapu, R., Mukkamala, R.R., Hussain, A., Flesch, B.: Social set analysis: a set theoretical approach to big data analytics. IEEE Access **4**, 2542–2571 (2016). https://doi.org/10.1109/ACCESS.2016.2559584
3. Colombo, P., Ferrari, E.: Access control in the era of big data: state of the art and research directions. In: Proceedings of the 23rd ACM on Symposium on Access Control Models and Technologies – SACMAT 2018, pp 185–192. ACM Press, New York, NY, USA (2018)
4. Bjurstrom, S.: Sentiment analysis methodology for social web intelligence. In: Proceedings of the Twenty-first Americas Conference on Information Systems. Association for Information Systems, Puerto Rico, pp 1–12 (2015)
5. Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C.: Social media analytics – challenges in topic discovery, data collection, and data preparation. Int. J. Inf. Manage. **39**, 156–168 (2018). https://doi.org/10.1016/j.ijinfomgt.2017.12.002
6. Feng, L., Chiam, Y.K., Lo, S.K.: Text-mining techniques and tools for systematic literature reviews: a systematic literature review. In: 2017 24th Asia-Pacific Software Engineering Conference (APSEC), pp 41–50. IEEE (2017)
7. Lorentzen, D.G.: Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. Scientometrics **99**, 409–445 (2014). https://doi.org/10.1007/s11192-013-1227-x
8. Sisodia, D.S., Reddy, N.R.: Sentiment analysis of prospective buyers of mega online sale using tweets. In: International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017, pp. 2734–2739 (2018). https://doi.org/10.1109/ICPCSI.2017.8392217
9. Boulos, M.N.K., Sanfilippo, A.P., Corley, C.D., Wheeler, S.: Social web mining and exploitation for serious applications: technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. Comput. Methods Programs Biomed. **100**, 16–23 (2010). https://doi.org/10.1016/j.cmpb.2010.02.007

10. de Carvalho, V.D.H., Costa, A.P.C.S.: Social web mining as a tool to support public security sentiment analysis. In: Freitas, P.S., Dargam, F., Ribeiro, R., et al. (eds.) 5th International Conference on Decision Support System Technology, pp. 164–169. EURO Working Group on Decision Support Systems, Funchal (2019)
11. Gerber, M.S.: Predicting crime using Twitter and kernel density estimation. Decis. Support Syst. **61**, 115–125 (2014). https://doi.org/10.1016/j.dss.2014.02.003
12. Nepomuceno, T.C.C., Costa, A.P.C.S.: Spatial visualization on patterns of disaggregate robberies. Oper. Res. (2019). https://doi.org/10.1007/s12351-019-00479-z
13. Pereira, D.V.S., Mota, C.M.M., Andresen, M.A.: The homicide drop in Recife, Brazil: a study of crime concentrations and spatial patterns. Homicide Stud. **21**, 21–38 (2017). https://doi.org/10.1177/1088767916634405
14. Henriques de Gusmão, A.P., Aragão Pereira, R.M., Silva, M.M., da Costa Borba, B.F.: The use of a decision support system to aid a location problem regarding a public security facility. In: Freitas, P.S.A., Dargam, F., Moreno, J.M. (eds.) EmC-ICDSST 2019. LNBIP, vol. 348, pp. 15–27. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-18819-1_2
15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**, 1–135 (2008). https://doi.org/10.1561/1500000011
16. Kharrat, S., Kchaou, S.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**, 267–307 (2007)
17. Brum, H.B., Das Graças Volpe Nunes, M.: Building a sentiment corpus of tweets in Brazilian Portuguese. In: LREC 2018 - 11th International Conference on Language Resources and Evaluation, pp. 4167–4172 (2019)
18. Chathuranga, J., Ediriweera, S., Hasantha, R., et al.: Annotating opinions and opinion targets in student course feedback. In: LREC 2018 - 11th International Conference on Language Resources and Evaluation, pp. 2684–2688 (2019)
19. Turchi, M., Negri, M.: Automatic annotation of machine translation datasets with binary quality judgements. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, pp. 1788–1792 (2014)
20. Win, S.S.M., Aung, T.N.: Automated text annotation for social media data during natural disasters. Adv. Sci. Technol. Eng. Syst. **3**, 119–127 (2018). https://doi.org/10.25046/aj030214
21. Walkowiak, T., Gniewkowski, M.: Distance measures for clustering of documents in a topic space. Adv. Intell. Syst. Comput. **987**, 544–552 (2020). https://doi.org/10.1007/978-3-030-19501-4_54
22. Cook, P., Brinton, L.J.: Building and evaluating web corpora representing national varieties of English. Lang. Resour. Eval. **51**, 643–662 (2017). https://doi.org/10.1007/s10579-016-9378-z
23. Hovy, E., Lavid, J.: Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. Int. J. Transl. **22**, 13–36 (2010)
24. Baccouche, A., Garcia-Zapirain, B., Elmaghraby, A.: Annotation technique for health-related tweets sentiment analysis. In: 2018 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2018, pp. 382–387 (2019). https://doi.org/10.1109/ISSPIT.2018.8642685
25. Zhang, H., Gan, W., Jiang, B.: Machine learning and lexicon based methods for sentiment classification: a survey. In: 2014 11th Web Information System and Application Conference (WISA). IEEE, New York, NY, USA, pp 262–265 (2014)
26. Neogi, P.P.G., Das, A.K., Goswami, S., Mustafi, J.: Topic modeling for text classification. In: Mandal, J.K., Bhattacharya, D. (eds.) Emerging Technology in Modelling and Graphics. AISC, vol. 937, pp. 395–407. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-7403-6_36
27. Dahal, B., Kumar, S.A.P., Li, Z.: Topic modeling and sentiment analysis of global climate change tweets. Soc. Netw. Anal. Min. **9**, 1–20 (2019). https://doi.org/10.1007/s13278-019-0568-8

28. Cunningham-Nelson, S., Baktashmotlagh, M., Boles, W.: Visualizing student opinion through text analysis. IEEE Trans. Educ. **62**, 305–311 (2019). https://doi.org/10.1109/TE.2019.2924385
29. Groß-Klußmann, A., König, S., Ebner, M.: Buzzwords build momentum: global financial twitter sentiment and the aggregate stock market. Expert Syst. Appl. **136**, 171–186 (2019). https://doi.org/10.1016/j.eswa.2019.06.027
30. Srinivasan, B., Mohan Kumar, K.: Flock the similar users of twitter by using latent Dirichlet allocation. Int. J. Sci. Technol. Res. **8**, 1421–1425 (2019)
31. Aggarwal, C.C.: Machine learning for text. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73531-3
32. Blei, D., Carin, L., Dunson, D.: Probabilistic topic models. IEEE Signal Process. Mag. **27**, 55–65 (2010). https://doi.org/10.1109/MSP.2010.938079
33. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl.-Based Syst. **89**, 14–46 (2015). https://doi.org/10.1016/j.knosys.2015.06.015
34. Yang, P., Chen, Y.: A survey on sentiment analysis by using machine learning methods. In: 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp 117–121. IEEE (2017)
35. Asghar, M.Z., Kundi, F.M., Ahmad, S., et al.: T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. Expert Syst. **35**, 1–19 (2018). https://doi.org/10.1111/exsy.12233
36. Khan, F.H., Bashir, S., Qamar, U.: TOM: Twitter opinion mining framework using hybrid classification scheme. Decis. Support Syst. **57**, 245–257 (2014). https://doi.org/10.1016/j.dss.2013.09.004
37. De Arruda, G.D., Roman, N.T., Monteiro, A.M.: An Annotated Corpus for Sentiment Analysis in Political News, pp. 101–110 (2015)
38. dos Santos, H.D.P., Woloszyn, V., Vieira, R., Blogset, B.R.: A Brazilian Portuguese blog corpus. In: LREC 2018 11th International Conference on Language Resources and Evaluation, pp. 661–664 (2019)
39. Freitas, C., Motta, E., Milidiú, R.L., César, J.: Sparkling Vampire… LOL! Annotating opinions in a book review corpus. In: Aluísio, S., Tagnin, S.E.O. (eds.) New Language Technologies and Linguistic Research: A Two-Way Road, pp. 128–146. Cambridge Scholars Publishing, Newcastle upon Tyne (2013)
40. de Souza, K.F., Pereira, M.H.R., Dalip, D.H.: UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro. Abakós **5**, 79 (2017). https://doi.org/10.5752/p.2316-9451.2017v5n2p79
41. Rosa, R.L., Rodriguez, D.Z., Bressan, G.: SentiMeter-Br: A new social web analysis metric to discover consumers' sentiment. In: Proceedings of the International Symposium Consumer Electronics, ISCE, pp. 153–154 (2013). https://doi.org/10.1109/ISCE.2013.6570158
42. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol (2009). https://www.nltk.org/
43. Reinoso, G., Farooq, B., Forum, C.T.R.: Urban pulse analysis using big data. In: Canadian Transportation Research Forum 50th Annual Conference. Transportation Association of Canada (TAC), Montreal, p. 16 (2015)
44. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)