Sensitive Keyword Extraction Based on Cyber Keywords and LDA in Twitter to Avoid Regrets

R. Geetha and S. Karthika

Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
geethkaajal@gmail.com, skarthika@ssn.edu.in

Abstract. Twitter is the most popular social platform where common people reflect their personal, political and business views that obliquely build an active online repository. The data presented by users on social networking sites are usually composed of sensitive or private data that is highly potential for cyber threats. The most frequently presented sensitive private data is analyzed by collecting real-time tweets based on benchmarked cyber-keywords under personal, professional and health categories. This research work aims to generate a Topic Keyword Extractor by adapting the Automatic Acronym - Abbreviation Replacer which is specially developed for social media short texts. The feature space is modeled using the Latent Dirichlet Allocation technique to discover topics for each cyber-keyword. The user's context and intentions are preserved by replacing the internet jargon and abbreviations. The originality of this research work lies in identifying sensitive keywords that reveal Tweeter's Personally Identifiable Information through the novel Topic Keyword Extractor. The potential sensitive topics in which the social media users frequently exhibit personal information and unintended information disclosures are discovered for the benchmarked cyber-keywords by adapting the proposed qualitative topic-wise keyword distribution approach. This experiment analyzed cyber-keywords and the identified sensitive topic keywords as bi-grams to predict the most common sensitive information leaks happening in Twitter. The results showed that the most frequently discussed sensitive topic was 'weight loss' with the cyber-keyword 'weight' of the health tweet category.

Keywords: Twitter · Cyber-keywords · Privacy leaks · Regrets · Social media

1 Introduction

One of the most popular social networking media, Twitter, is now emerging as a platform to share what users feel, experience and comment both on personal and nation's current affairs with 280 characters per tweet. Almost all tweets contain some personal value when tweeted or re-tweeted by the users. Though sharing personal information in social media gives pleasure, the extent to which the message travels have no bounds or could not be limited as the user believes. Also, the users feel free to publish their opinion and

Published by Springer Nature Switzerland AG 2020
A. Chandrabose et al. (Eds.): ICCIDS 2020, IFIP AICT 578, pp. 59–70, 2020.
https://doi.org/10.1007/978-3-030-63467-4_5

thoughts online than sharing it in-person because this networking platform provides a wide range of audience which is both undefined and unlimited.

Gathering private information in this digital era has become much easier because of the advancements in modern communication technologies. Tommasel and Godoy (2016) stated that the access to social media posts are also widely open to every user irrespective of their age, location, expertise and professional community boundaries which leads to leaks in Personally Identifiable Information (PII) either knowingly or unknowingly. The publicly available personal information is at high risk of potentially being misused by cyber criminals eventually resulting in diverse effects like identity theft, child trafficking, business loss and professional black marks. A lot of issues related to privacy are arising in social media and plenty of privacy protection technologies are designed for users which tend to be independent and application specific. Such privacy protection mechanisms usually aim at protecting the information and control information disclosure to unintended audiences.

In-spite of all, the protection mechanisms available for both online and offline, there is a drastic increase in cyber-crimes. Therefore, some kind of prevention strategies should be adopted in the user's side to withstand and overcome such cyber-crimes. One such strategy is proposed in this research work by highlighting the importance of text that users post on social media. Users should be ethically aware of what content is presented to whom, who all are accessible to what all contents and what contents should be excluded in the post. The authors define the sensitivity of a user tweet based on many existing researches and user experiences.

Sensitivity

In context with this research work, sensitivity in a social media post, say Twitter, should be broadly categorized into three domains namely personal, professional and health-oriented information shared publicly in a social media platform. The sensitivity in a user message can be referred to "*a private identity or incriminating information disclosed by either the data owner itself or others without the consent or knowledge of the data owner*". Here, the data owner can signify to a person or organization or community who can be identified or traced through the sensitive content posted by the social media user.

Hence the task of sensitive text classification is integrated for identifying the tweet sets, text categories and topics, taking text analytics to the next level as described by Hu et al. (2012). The objective of the research work is to formulate better tweet pre-processing rules, impose exact words for internet jargons and abbreviations in tweets, identify the sensitive keywords for personal, professional and health related tweets by modeling a framework called Sensitive Topic Keyword Extractor.

The major contributions of this research work are listed as follows:

1. To build an Automatic Acronym-Abbreviation Replacer for twitter community
2. To determine what is sensitive in personal, professional and health tweet domains
3. To generate unique topic models using LDA for personal, professional and health tweet domains
4. To investigate sensitive keywords that co-occur with the cyber-keywords in personal, professional and health tweet domains

2 Related Work

2.1 Twitter and Privacy Concerns

Lu et al. (2015) defined two privacy attributes namely confidentiality and universality with a set of 53 keywords. The privacy and security levels of Internet of Things (IoT) and Cyber Physical Systems (CPS) are classified by the Privacy Information Security Classification (PISC Model). Various challenges that are related to detecting privacy related information from unstructured texts are discussed by Sleeper et al. (2013) using natural language processing, ontology and machine learning approaches. It also deals with the problem of identifying the user's perception, domain specificity, context dependence, privacy sensitivity classification, data linkages in the social media messages. Wang et al. (2011) conducted a survey for capturing regrets that were caused by misunderstanding, misuse, misjudging of online posts and discovered the mechanisms preferred by the users to overcome the regrets such as apology, decline or delay the request, etc.

2.2 Twitter Topic Identification

The research work carried out by Aphinyanaphongs et al. (2014) presented a feasibility study of identifying and inferring the usage patterns and most frequent places where users posts alcohol related tweets using various classification techniques such as support vector machines, Naive Bayes, random forest and Bayesian logistic regression. Since natural language processing tools cannot be directly applied to social media texts, it is necessary to modify the extracted data so that it resembles standard texts. Baldwin et al. (2013) used data from various sources like Twitter, YouTube and Facebook comments, forums, blogs, Wikipedia and British National Corpus (BNC) to build corpus and performed lexical analysis, grammatical checks, language modeling and corpus similarity. A novel, unsupervised approach for detecting topic on twitter by using Formal Concept Analysis (FCA) is presented by Cigarrán et al. (2016). To handle the adverse effects of feature sparsity encountered by Tesfay et al. (2016) and curse of dimensionality in short text message classification, Lei et al. (2012) introduced the concept of boosting called LDABoost designed with the Naïve Bayes algorithm as a weak classifier. To improve the topic learning in social media messages, Mehrotra et al. (2013) proposed a novel approach of pooling and aggregating tweets in the tweet pre-processing step carried out for LDA. Whereas to learn the topic and behavior interests of the users, Qiu et al. (2013) introduced Behavior LDA (B-LDA) model using Twitter messages.

3 Methodology

The proposed Sensitive Topic Keyword Extractor analyses large volume of user generated tweets collected using the Twitter Streaming API for 68 cyber-keywords grouped under three categories namely personal, professional and health domains as derived from the findings of Lu et al. (2015) and Mao et al. (2011). The raw tweets are cleaned and prepared with pre-processing rules resulting with tweets suitable for text classification and analysis. The data annotation process for identifying sensitive tweets was performed using the most popular and reliable crowd sourcing platform called Amazon Mechanical Turk (AMT).

The LDA, being the most powerful probabilistic model for text classification is opted for identifying the topic keywords for each tweet category. The objective of the Sensitive Topic Keyword Extractor is to identify the next level of sensitive keywords that would cause potential vulnerability when used with the benchmarked cyber-keywords. The sensitive keywords that occur with the cyber-keywords are extracted using the Variational Expectation Maximization (VEM), VEM-Fixed and Gibbs model by filtering the non-cyber keywords present in keyword extraction process for various topics. Identifying such sensitive keywords will help users in realizing the potential threat caused by posting information related to those keywords. A detailed architecture of the proposed system, Sensitive Topic Keyword Extractor is represented in Fig. 1.

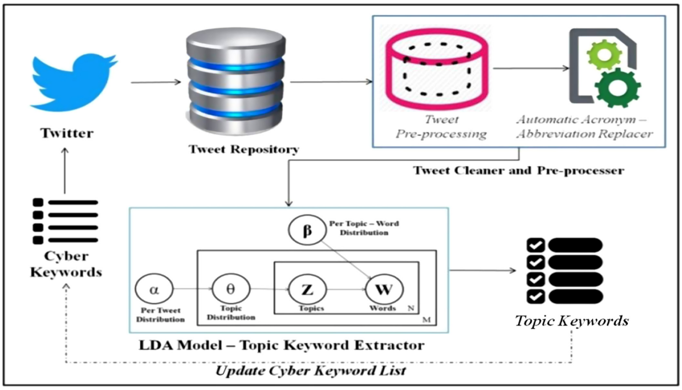

Twitter
Tweet Repository
Tweet Pre-processing
Automatic Acronym – Abbreviation Replacer
Tweet Cleaner and Pre-processer
Cyber Keywords
β
Per Topic – Word Distribution
α
Per Tweet Distribution
θ
Topic Distribution
Z
Topics
W
Words
N
M
LDA Model – Topic Keyword Extractor
Topic Keywords
Update Cyber Keyword List

Fig. 1. Architecture of the proposed system – Sensitive Topic Keyword Extractor

3.1 Tweet Cleaning and Pre-processing

The tweets are short text messages posted by users without any formal template or sentence structure. The tweets tend to have short forms, internet jargons, abbreviations, symbols, numbers, date and time, email, URLs and other non-dictionary words as mentioned by Sproat et al. (2001). Therefore, it is necessary to wisely handle such inappropriate content and perform data transformation operations as listed in Table 1. During the process of cleanup, it is necessary to replace acronyms with corresponding abbreviations. Hence a dedicated list of acronyms and abbreviations is prepared with respect to twitter slang and internet jargons which will be used to handle the automatic replacement of short forms in the tweet.

3.2 Topic Keyword Detector Using LDA Model

The pre-processed tweets are taken to construct the term-document matrix which maps the words to its importance in the tweet and the overall tweets set under each category say

Table 1. A list of pre-processing rules used by Sensitive Topic Keyword Detector.

Pre-Processing Rule
Convert t to lowercase.
Remove"@" in user mentions and "#" in hashtags.
Replace URLs → 'url', date → "ddmmyy", time → "hhmm", phone numbers → "ph_no".
Remove stopword and perform stemming for each word.
Remove symbols, punctuations, numbers, unwanted whitespaces and new lines.
Replace acronyms with corresponding abbreviations and internet jargons.

personal, professional and health. There are basically three features considered for text classification namely Feature Frequency (FF), Term Frequency – Inverse Document Frequency (TF-IDF), Feature Presence (FP). The LDA model, an unsupervised probabilistic model generally combines topic, context and word through probability by focusing only words. The topics and keywords are extracted based on two parameters namely per document distribution and per topic word distribution. The LDA model generally has three steps in detecting topics. Firstly, the model identifies the term distribution β for each topic specified. Secondly, the proportion of each topic is computed for each tweet. Thirdly, for every word w_i in tweet t, a multinomial function is selected for each topic associated with that word; a word w_i is selected from that multinomial function based on the probability distribution for each topic. The domain-wise sensitive keywords are extracted by setting a threshold value $Ø_{tc}$ for each topic in a tweet category, the keywords are selected based on the conditional probability values. The results showed a set of well associated domain-wise sensitive keywords for all three models taken into consideration namely VEM, VEM_Fixed and Gibbs model used for topic learning and detection. The VEM is similar to Expectation Maximization technique with the difference of iteratively performing the expectation step until the likelihood becomes computationally intractable.

4 Results and Discussions

4.1 Dataset

The benchmarked cyber-keywords pre-defined by Lu et al. (2015) comprised of 23 keywords for personal topics, 22 keywords for professional topic and 7 for health topics. Since there are only 7 cyber-keywords related to health, an additional set of 16 health-related cyber-keywords were added by filtering out the most frequently used health keywords in internet as mentioned by Geetha et al. (2019). Finally, a list of 68 cyber-keywords are defined and used for tweet collection. The authors avoided manual inclusion of keywords in the cyber-keyword list due to the factor that it might lead in compromising the quality and relevance of data collected for sensitive data analysis. A large volume of real time tweets were collected that counted to 800900 tweets. This collection contributed about 36% of personal tweets, 32% of professional tweets and 32% of health-related tweets. As depicted in Fig. 2, on building the corpus data for analysis, the authors chose to consider only the original tweets by eliminating the retweets from the data repository. The tweet corpus built comprised of 35%, 33% and 32% of the total tweets collected for personal, professional and health related tweets.

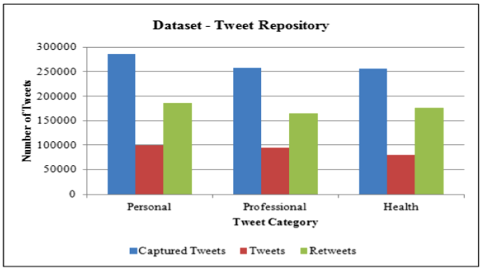


Fig. 2. The statistical representation of tweets collected for three tweet categories.

Dataset Annotation

The original tweets are further considered for annotation process as a Human Intelligence Task (HIT) that is performed by the AMT domain expert workers. The authors presented a detailed documentation composed of the following three itineraries to the AMT domain expert workers.

(i) What is sensitivity in a social media text?
(ii) A ruleset for identifying the presence of sensitive private data or PII.
(iii) A set of gold data with sensitive and insensitive tweets in three tweet domains.

The set of gold data that were submitted to the AMT workers was the driving factor to arrive at a clear understanding and converge the rules and assumptions on sensitivity with the given tweets to identify the presence of PII. A sample gold data submitted to AMT domain expert workers is presented in Table 2.

On submission of the dataset, the authors received annotation results from 670 AMT works for 54176 tweets. The authors selected 153 out of 670 AMT domain experts by considering their knowledge in social media, turnaround time and consistency in annotating 50 tweets per session. The annotation process was completely assisted by the authors in case of any queries from the AMT domain expert workers. The annotation results submitted by the AMT domain expert workers were evaluated by applying the approach of multiple grading and gold standard evaluation. The tweets after the evaluation process is preprocessed with a set of standard data cleaning steps from Table 1 and stored in the tweet corpus as described in the Table 3.

4.2 Sensitive Topic Keyword Extractor Using LDA Model

The pre-processed tweets are taken for corpus building which in-turn is given for generating the bag-of-words by forming term document matrix. For text categorization,

Table 2. A sample of gold data submitted to the AMT domian expert workers.

Tweet domain	Gold set data	Annotation
Personal tweet	I got into a car accident last night and I think I'm still in shock	Sensitive
	It's getting easy for criminals to get anyone's mobile call record illegally. @hydcitypolice @CPHydCity @KTRTRS @amjedmbt @asadowaisi	Insensitive
Professional tweet	@AmericanAir Hi. I purchased a ticket online, the money was taken from my bank account, but I have no record locator or conf. email. Help!!	Sensitive
	These cards are submitted like insurance in the pharmacy. So the company now has your name, address, DOB, phone number, and med history.	Insensitive
Health tweet	Having RA has been an absolute test for me but I'm learning to live life w/this disease & it's getting more manageable day by day ??	Sensitive
	Diabetes-related eye disease often has no symptoms until it reaches an advanced stage #WorldSightDayAU	Insensitive

Table 3. Transformation of a sample tweet in data cleaning and pre-processing stage.

Sample tweet	FYI: For the next 7 days I will not have access to SMS or mobile phone. Please use email, WhatsApp, Viber, … https://t.co/IDmYVyJnGy
Pre-processed tweet	For your information for the next 7 days i will not have access to short message service or mobile phone please use email whatsapp viber url
Tweet in corpus	Information next day access short message service mobile phone email whatsapp viber

feature selection is based on the selection of appropriate words by using dictionaries and NLP standards. Thus, this paper uses optimal text replacement strategies for building better feature space by using the proposed Automatic Acronym - Abbreviation Replacer (AAAR) specially designed for twitter. The AAAR was built with 1200 internet jargons and twitter slangs that are most popularly used by users and frequently encountered in our dataset. The AAAR is subject to change and update in accordance with the evolving practices and trends of online social networks.

The generated matrix is applied to three different models of LDA namely VEM, VEM-Fixed and Gibbs model to find topic keywords provided the k number of keywords to be found under n topics. The three category of tweet sets are taken individually and applied for LDA with k = 10; n = 5 and k = 20 and n = 10. Therefore, it generates 5 sets of 10 topic keywords for 5 topics and 10 sets of 20 topic keywords for 10 topics for all three LDA models. The sensitive keywords are identified by filtering out the cyber-keywords from the generated topic keywords.

Table 4 projects the overview of the identified sensitive keywords under each tweet category with cyber-keywords. The α value, parameter for finding the per topic distribution in tweet category denoted the confidence of the topic assignment. It can be inferred that, higher the value of α, more even the distribution of topics over the tweets. The lower the value of α, higher the percentage of tweets been assigned to one particular topic.

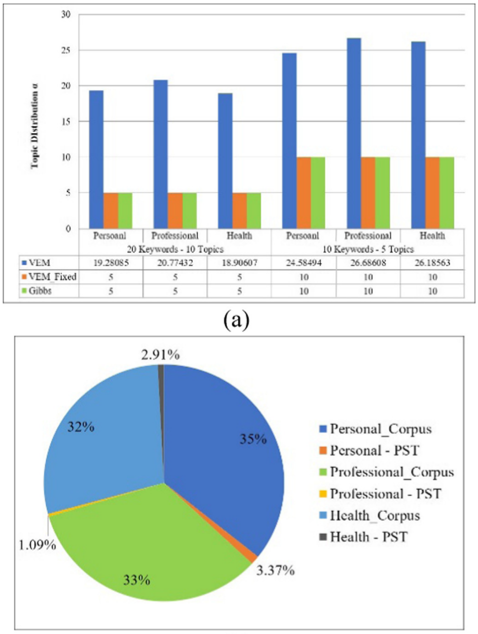
Table 4. A list of bench-marked cyber-keywords and identified sensitive keywords by the proposed strategy for personal, professional and health related tweets.

Tweet category	No. of cyber-keywords	Sample cyber-keywords	Identified sensitive keywords
Personal	23	car, chat, family, hobby, marriage, phone number. religion, spouse	friend, today, message, happy, time, god, live, stay, nice, great, work, read, application, hope
Professional	22	company address, insurance, investment, bank account, password, credit score, passport	twitter, real, money, order, private, trump, pay, state, incoming, market, debt. business, tax, team
Health	23	birth, blood type, height, disease, vaccine, therapy	food, body, world, home, mom, sinus, cancer, care, women, water, flu, sex

Therefore, Fig. 3 shows the topic distribution values achieved for all three tweet categories with three LDA models. It can be inferred from the topic modeling results that as the number of topic increases, the α value reduces which denotes the uneven distribution of topic keywords. The VEM model performs better than other two models with better variations in topic distribution for all three tweet categories as depicted in Fig. 4. The topic distribution for personal tweets with 20 keywords under 10 topics was 19.28085 whereas for 10 keywords with 5 topics was 24.58494. This clearly shows that as the number of topics increases, the keywords are more widely spread. The same behavior can be observed in all three tweet categories. This effect is due to the high sparsity that exists in social media messages which should be handled by adapting feature selection mechanisms.

The optimal model for sensitive keyword extraction was identified to be LDA with VEM model using TF-IDF for feature extraction with higher F1 score as shown in Fig. 4. The topic-wise keyword distribution achieved for each tweet category and a threshold $Ø_{tc} = 0.3$ determined by evaluating the model with varying iterations $nIter = \{10, 20, 30, 40, 50\}$ derived a set of 93 sensitive keywords.

The authors combined all cyber-keywords and identified sensitive keywords into bigrams from which the most frequently discussed sensitive topic would be identified as shown in Fig. 5. It can be inferred from the bi-gram analysis that in personal tweet category, the cyber-keyword ‘phone’ and sensitive keyword ‘love’ were frequently detected



(a)

(b)

Fig. 3. **a.** The topic distribution parameter - α value achieved for three tweet category for VEM, VEM_Fixed and Gibbs Model to extract 20 keywords under 10 topics and 10 keywords under 5 topics. **b.** The proportion of Potentially Sensitive Tweets (PST) in the tweet corpus for three categories of tweet.

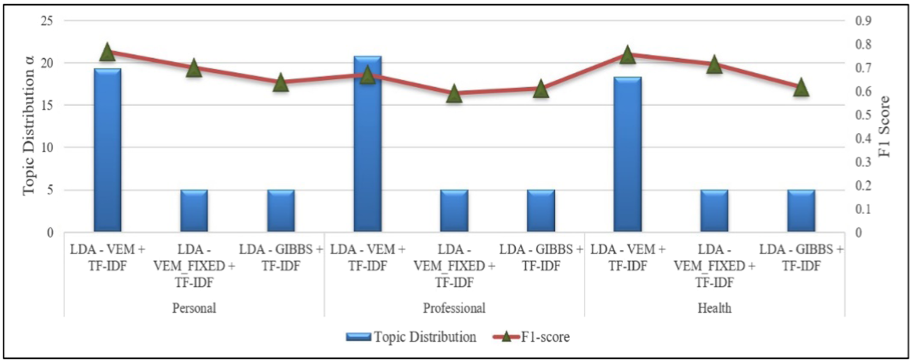


Fig. 4. The topic-wise keyword distribution (10 topics – 20 keywords) and F1 score achieved through various sampling models of LDA for three tweet categories.

to be occurring together. Similarly, the professional tweet category has 'account' and 'money' and health tweet category has 'weight' and 'loss' as the most frequently occurring cyber-keyword and sensitive keyword respectively. These co-occurring keywords are subject to vary when the analysis is performed with future user tweets in the same tweet domains.

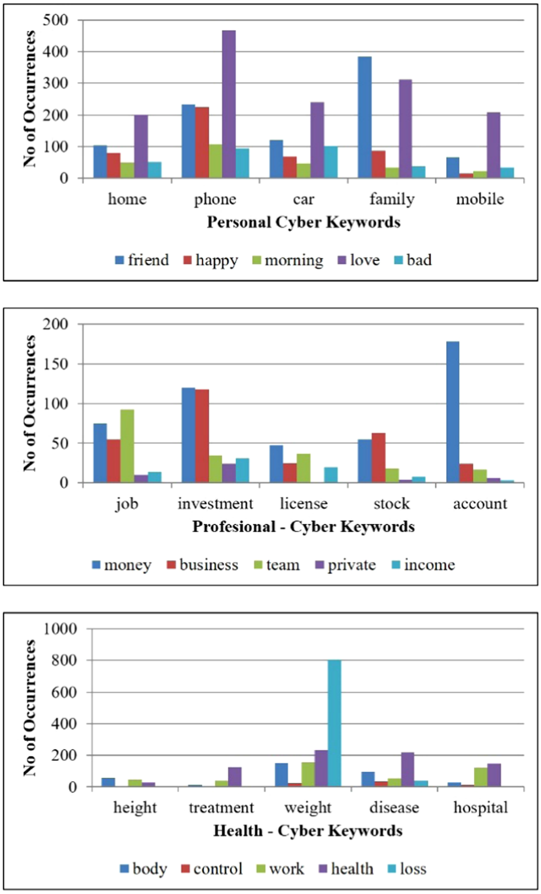


Fig. 5. The Topic-wise keyword distribution obtained for benchmarked cyber-keywords and sensitive keywords of three tweet categories modeled as bi-grams.

The existing researches that employ LDA in identifying topics has many scopes for sensitivity topic identification. Table 5 presents the performance analysis of most relevant LDA models by Gunawan et al. (2018), Zou and Song (2016) against the proposed method using the F1 measure thereby highlighting the suitability and uniqueness of the proposed approach for sensitive data identification in Tweets. Therefore, from the observed research results, the authors concluded that users should be cautious when using these cyber and sensitive keywords together because they are believed to have high potential in revealing personally identifiable information.

Table 5. Comparison of results for the sensitive data for various LDA models.

Dataset		
8,00,900 Tweets in three tweet domains		
Personal	2,86,249	
Professional	2,58,493	
Health	2,56,158	
Performance Comparison - F1 measure		
Twitter LDA – GIBBS (Gunawan et al. 2018)		
Personal	Professional	Health
0.642	0.59	0.667
LDA + TF-IDF (Zou et al. 2016)		
Personal	Professional	Health
0.695	0.653	0.638
Proposed Method		
LDA – VEM + TF-IDF		
Personal	*Professional*	*Health*
0.7581	*0.6682*	*0.7558*
LDA – VEM-Fixed + TF-IDF		
Personal	Professional	Health
0.6998	0.5891	0.7145

5 Conclusion

This research work discovers sensitive keywords in addition to the benchmarked cyber-keywords that are more frequently used along the cyber-keywords in user tweets. An enhanced data cleaning mechanism is implemented in replacing the abbreviations and internet jargons by building an Automated Acronym - Abbreviation Replacer module. The major contribution of this research work is building a framework called Sensitive Topic Keyword Extractor which the identified of 93 sensitive keywords that are potentially sensitive which when used with the benchmarked 68 cyber-keywords. Therefore, the sensitive keywords are also considered as critical or sensitive keywords and will be used in tweet collection and tweet repository building for the future works which will result in having a set of 161 query keywords for tweet extraction.

References

Tommasel, A., Godoy, D.: Short-text feature construction and selection in social media data: a survey. Artif. Intell. Rev. **49**(3), 301–338 (2016). https://doi.org/10.1007/s10462-016-9528-0

Hu, X., Liu, H.: Text analytics in social media. Min. Text Data, 385–414 (2012)

Lu, X., Qu, Z., Li, Q., Hui, P.: Privacy information security classification for internet of things based on internet data. Int. J. Distrib. Sens. Netw. **11**(8), 932–941 (2015)

Sleeper, M., et al.: I read my Twitter the next morning and was astonished: a conversational perspective on Twitter regrets. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3277–3286 (2013)

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P.G., Cranor, L.F.: I regretted the minute I pressed share: a qualitative study of regrets on Facebook. In: Proceedings of the Seventh Symposium on Usable Privacy and Security, p. 10. ACM (2011)

Aphinyanaphongs, Y., Ray, B., Statnikov, A., Krebs, P.: Text classification for automatic detection of alcohol use-related tweets: a feasibility study. In: 2014 IEEE 15th International Conference on Information Reuse and Integration (IRI), pp. 93–97 (2014)

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L.: How noisy social media text, how diffrnt social media sources? In: IJCNLP, pp. 356–364 (2013)

Cigarrán, J., Castellanos, Á., García-Serrano, A.: A step forward for topic detection in Twitter: an FCA-based approach. Expert Syst. Appl. **57**, 21–36 (2016)

Tesfay, W.B., Serna, J., Pape, S.: Challenges in detecting privacy revealing information in unstructured text. In: PrivOn@ ISWC (2016)

Lei, L., Qiao, G., Qimin, C., Qitao, L.: LDA boost classification: boosting by topics. EURASIP J. Adv. Signal Process., 233 (2012)

Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 889–892 (2013)

Qiu, M., Zhu, F., Jiang, J.: It is not just what we say, but how we say them: LDA-based behavior-topic model. In: Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 794–802. Society for Industrial and Applied Mathematics (2013)

Mao, H., Shuai, X., Kapadia, A.: Loose tweets: an analysis of privacy leaks on Twitter. In: Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society, 1–12 (2011)

Sproat, R., Black, A.W., Chen, S., Kumar, S., Ostendorf, M., Richards, C.: Normalization of non-standard words. Comput. Speech Lang. **15**(3), 287–333 (2001)

Gunawan, D., Rahmat, R.F., Putra, A., Pasha, M.F.: Filtering spam text messages by using Twitter-LDA algorithm. In: 2018 IEEE International Conference on Communication, Networks and Satellite (Comnetsat), pp. 1–6. IEEE, November 2018

Zou, L., Song, W.W.: LDA-TM: a two-step approach to twitter topic data clustering. In: 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 342–347. IEEE, July 2016

Geetha, R., Karthika, S., Pavithra, N., Preethi, V.: Tweedle: sensitivity check in health-related social short texts based on regret theory. Procedia Comput. Sci. **165**, 663–675 (2019)