Violent Vocabulary Extraction Methodology: Application to the Radicalism Detection on Social Media

Amal Rekik[1,2](✉), Salma Jamoussi[1,2], and Abdelmajid Ben Hamadou[1,2]

[1] Multimedia InfoRmation Systems and Advanced Computing Laboratory MIRACL, University of Sfax, Sfax, Tunisia
rekik.amal91@gmail.com, jamoussi@gmail.com, abdelmajid.benhamadou@isimsf.rnu.tn
[2] Digital Research Center of Sfax DRCS, 3021 Sfax, Tunisia

Abstract. Nowadays, social networks have become powerful mediums of communication providing information, learning and entertainment. Unfortunately, these platforms can be sorely manipulated by vicious users sharing malicious contents. Therefore, the process of mining and analyzing such published suspicious content is a considerably challenging task that serves to fight against the online radicalization. For this purpose, we propose, in this paper, a new methodology for extracting and analyzing violent vocabulary shared on social networks with the exploration a set of natural language processing and data mining techniques. Our method relies mainly on extracting a set of profiles judged by a domain expert as extremist and non-extremist' users. Then, we focus on their shared textual content in order to detect malicious vocabulary published within the radical context as well as their violence' degrees. Finally, in order to evaluate the performance of our method, we resort to an expert who verifies the final list of the extracted vocabulary annotated by our method. Thus, the given results show its effectiveness as well as its efficiency.

Keywords: Violent vocabulary · Vocabulary extraction · Terrorist user · Radicalism detection · Social network analysis

1 Introduction

Today, the religious radicalization becomes one of the most furious peril threatening the public security around the whole world. Indeed, the term radicalization is implied as a process through which individuals evolve towards extremism by either practicing, promoting or defending violence to achieve their goals. For this purpose, this trend should be seen in light especially with the emergency of the enormous impact of social networks which ensure the swift development of online radicalism. In fact, radical organizations rely heavily on social networks to promote extremism and share suspicious content which can be either open and transparent or under cover and coded. As a result, social networks have become a gateway for extremists and a starting point of radicalization playing the role of facilitators and amplifiers within the community. On

N. T. Nguyen et al. (Eds.): ICCCI 2019, LNAI 11684, pp. 97–109, 2019.
https://doi.org/10.1007/978-3-030-28374-2_9

the other hand, extremists on these networks are frequently exploring a terribly evil, violent and vindictive vocabulary in order to propagate their vicious ideas. Therefore, the process of mining and analyzing such published suspicious content on social networks is a considerably challenging task that raises several requirements to be addressed and serves to fight against the online radicalization. So, faced to this dangerous phenomenon, the prevention of the online radicalization has become a very indispensable action that frustrates recruitment and avoids transitions from excitement of emotions and ideological influence to the active participation by force and violence.

In this context, we propose in this paper a new methodology for violent vocabulary extraction in order to detect radicalism on social networks. Our method relies first on a set of social networks collected profiles annotated by a domain expert as extremist and non-extremist' users. Then, we focus precisely on their textual content in order to extract the vocabulary specific to both radical and non-radical contexts. This analyzed content is generally shared in the Arabic language which raise additional requirements to be respected in the context of data analyzes field. Finally, by using a set of natural language processing and data mining techniques, our methodology attempts to extract a fierce vocabulary weighted by their violence's degree. In fact, exploring our methodology to analyze malicious shared content and extract violent vocabulary specific to radical discourses leads to discover various extremists' profiles and thus detect radicalism on social networks.

The reminder of this paper is planned as follows: the next section is devoted for the related work. In Sect. 3, we go into details of our proposed methodology for the violent vocabulary extraction from social networks. Section 4 reports some statistics of the collected vocabulary and the experiments results relying on the expert observations. We end up with the conclusion and the future work in Sect. 5.

2 Related Work

The vocabulary extraction and the social network analysis are two of the most important tasks that interest researchers in the data mining field. In this context, several studies have been proposed in the literature [1–5]. So, this section surveys previous work in the vocabulary extraction field.

In [6], the author proposes a new principle for vocabulary selection dedicated for the theme detection task. This method is based mainly on the preposition that each theme is defined by its own vocabulary. Moreover, the author assumes that if the vocabulary selection method is the same as the word frequency method, some non-important frequent word will be selected to form the vocabulary. While, other non-frequent and relevant words will not be preserved. For this purpose, the proposed method consists first on evaluating the frequency measure for all the words of the theme's learning, for each given theme. Then, the author keeps only the words with the highest values. In this case, he obtains a vocabulary for each treated theme. Next, he performs the union of these obtained vocabularies by using the same size of each theme's vocabulary or using different sizes in order to form the final vocabulary.

In [7], authors employ the SVM applying the Random forest technique in order to select the vocabulary. In fact, the Random forest is a classification method that provides feature importance. The decision tree forest algorithm performs learning on multiple decision trees trained on slightly k different subsets of data. The k predictions of the variable of interest are stored for each original observation. Then, the prediction of the random forest is deducted as a simple majority vote through the Ensemble learning algorithm. Thus, important features are selected to form the final vocabulary.

In [8], author proposed an unsupervised algorithm to select task-specific Chinese words from a large general vocabulary. The proposed method is based on measuring the correlation of two adjacent character strings by calculating their mutual information (MI) [9]. Authors segmented the training data into word sequences based on a general vocabulary. Each word can be segmented into two adjacent character strings. Next, they computed the MI of each word by choosing the segmentation that provided a minimum MI value. If the MI of a word is greater than a predefined threshold, they consider that it matches the target task. Thus, this word belongs to the extracted vocabulary.

In [10], authors assume that the vocabulary used by each individual is a combination of two vocabularies which are: A topic-independent lecture vocabulary, that contains vocabulary common to spontaneous speech, and a topic dependent vocabulary. So, authors proposed a novel method for vocabulary selection to automatically adapt automatic speech recognition This method proceeds with the topic-independent lecture vocabulary, that consists of stop words and common words used in spontaneous lecture speech. Moreover, they select a topic-specific vocabulary for each lecture utilizing materials that are available before the lecture begins, like lecture slides. Using these documents, the proposed method collects automatically a large corpus of related documents from the World-Wide-Web. Finally, an active recognition vocabulary is selected using a feature-based word ranking computed using this corpus.

So, during this stage, we have arrived to deliver the weaknesses of the studied approaches. In fact, some methods do not occur in the language we are dealing with which is Arabic. Other methods neglect the semantic context during the extraction of vocabulary. Thus, all these methods are not designed exactly for our goals since we are mainly interested on the dangerous vocabulary extraction which is a highly sensitive field.

3 Proposed Methodology

Online religious radicalization on social networks is a malicious stir which is generally practiced by Arabic extremists. Thus, we are mainly interested in our methodology by the Arabic language analyzes in order to extract the violent vocabulary shared by radical communities. Meanwhile, analyzing Arabic language contents raises several requirements to be addressed and assists to make of our methodology a scoop in the data mining field. In addition, extracting such dangerous vocabulary serves to discover several extremists' users and thus extract radicalism on social networks. Our methodology contains mainly three major stages: (1) The suspicious data collection step from two social media sites which are Twitter and YouTube. (2) The frequent n-grams and itemsets extraction from the collected contents as well as their violence' degrees. These two primal steps are accomplished for both radical and non-radical communities. (3) The violent vocabulary extraction. Figure 1 represents the overall process of our methodology.

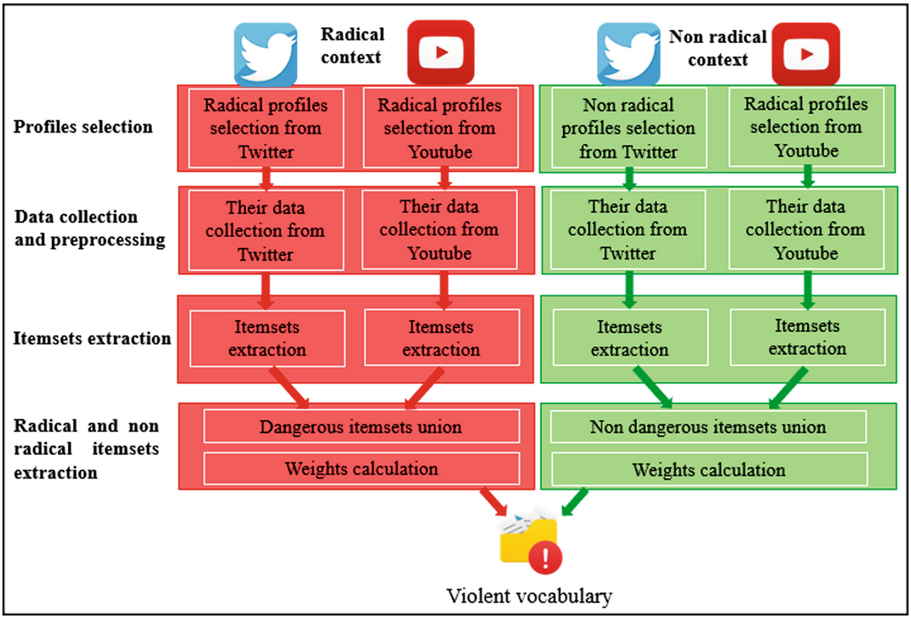
Radical context
Non radical context
Profiles selection
Data collection and preprocessing
Itemsets extraction
Radical and non radical itemsets extraction
Radical profiles selection from Twitter
Radical profiles selection from Youtube
Non radical profiles selection from Twitter
Radical profiles selection from Youtube
Their data collection from Twitter
Their data collection from Youtube
Their data collection from Twitter
Their data collection from Youtube
Itemsets extraction
Itemsets extraction
Itemsets extraction
Itemsets extraction
Dangerous itemsets union
Weights calculation
Non dangerous itemsets union
Weights calculation
Violent vocabulary

Fig. 1. The overall process of our proposed methodology

3.1 Data Collection and Preprocessing

The data collection step consists mainly on extracting textual content shared by both radical and non-radical communities on Twitter and YouTube. For this purpose, we start initially by collecting a set of extremists and non-extremist users from these two social media sites. To do so, we explore our data collection methodology proposed in [11]. This later consists first on preparing an incremented list of dangerous keywords which are judged by a domain expert such as Islamic state, Sharia, Jihad… Then, it excerpts Twitter users who talk about these suspicious keyword as well as the implicated users in terrorists' attacks. Nevertheless, the strategy adopted to collect data from YouTube consists on keeping users that are reactive towards videos of extremism and thus have a clear attitude to the radicalism. Then, the collected profiles are annotated by a domain expert in order to distinguish between radical and non-radical users. So, exploring this methodology, we conserve two communities in order to analyze their textual content: (1) the radical community containing extremist users and (2) the non-radical communities holding non-extremist users.

After this step, we proceed the data collection stage. So, we extract all the published textual data of each Twitter user belonging to each community. On the other hand, for every YouTube user of each community, we keep all his comments toward malicious videos as well as the titles and descriptions of his liked and shared videos.

Originally, the data extracted from these social media is reached by open sources obtained from the content generated by different users. Meanwhile, the results of the data mining method depend heavily on the quality of the data. For this reason, the preprocessing of the collected data is very crucial to achieve a performant data analyzes step. Therefore, during this stage, we perform the preprocessing of the collected data. To do so, we clean first all characters that are not in Arabic Language. Then we remove diacritics, punctuation marks, numbers and stop words from the Arabic textual data. Hence, we obtain two final datasets appropriated for radical and non-radical textual content which are prepared to be analyzed.

3.2 N-Grams and Itemsets Mining

After the data collection stage, we move to the extraction of the frequent n-grams and itemsets from the two collected preprocessed datasets. In fact, radical organizations as well as their used vocabulary differ from one social network to another. For this purpose, this step consists on analyzing the data of each community collected from Twitter and Youtube separately and then merge the obtained results. So, initially, we have separated each data shared by the radical community on Twitter and Youtube in n-grams with n <= 3. Similarly, we have represented each data shared by the non-radical community in n-grams with n <= 3. In fact, an n-gram is defined as a subsequence of n words constructed from a given sequence respecting the order. Some examples of these used n-grams are represented as follows:

Unigram (n=1): Example: "دولة (state)"," يوم (day)", …
Bigram (n=2): Example: "دولة إسلام (Islam state)"; "متحف باردو (Bardo museum)", …
Trigram (n=3): Example: "الإمارات العربية المتحدة" (United Arab Emirates), …

After the representation of the textual content shared by both radical and non-radical communities, we proceed the frequent n-grams and itemsets extraction step. Actually, an itemset is an association of n-grams occurring together in the collected data ignoring the order. Each n-gram has a support which is defined as its frequency in the collected dataset. Likewise, each itemset possesses a support corresponding to the frequency of simultaneous appearance of n-grams contained in the data set and which can be calculated as follow:

$$Support(itemset\ I) = \frac{NumberDataContaining(I)}{TotalNumberOfData}$$

Where NumberDataContaining(I) refers to the number of data composing the itemset I, and TotalNumberOfData is the size of the overall dataset.

Each n-gram or itemset is considered as frequent if its support is greater than or equal to a predefined threshold. Indeed, the frequent n-grams and itemsets extraction from data play one of the most essential role since it tries to find interesting patterns from databases, and thus constitutes the core stage of our methodology.

At this step, our approach explores the A-priori algorithm in order to extract the frequent n-grams and itemsets. In fact, the A-priori algorithm is a data mining algorithm designed in the field of learning association rules. It is used to recognize properties that come up frequently in the analyzed data.

In the dataset appropriated to the radical community, there are x elements corresponding to the tweets extracted from radical organizations on Twitter and Youtube also known as transactions. Each element is described by a set of attributes A = $\{a_i\}$ where each a_i corresponds to an item or an itemset.

So, we have extracted all the n-grams and itemsets and then calculated their supports to prune those which are non-frequent. Next, we keep as frequent n-grams and itemsets the union of the obtained frequent n-grams and itemsets from Twitter and Youtube. The final support of each obtained n-grams or itemsets is calculated as follow:

If the n-gram or the itemset is acquired from Twitter and not obtained from Youtube, then:

$$Support(itemset\ I) = Support(I)_{Twitter}$$

If the n-gram or the itemset is acquired from Youtube and not obtained from Twitter, then:

$$Support(itemset\ I) = Support(I)_{Youtube}$$

If the n-gram or the itemset is acquired from both Twitter and Youtube, then:

$$Support(itemset\ I) = \frac{Support(I)_{Twitter\,+}\, support(I)_{Youtube}}{2}$$

Where $\text{Support(I)}_{\text{Twitter}}$ is the support of the itemset I extracted from Twitter, and $\text{Support(I)}_{\text{Youtube}}$ is the support of the itemset I extracted from Youtube. After the computation of the support of each n-gram and itemset, we notice that these obtained supports are closed and scattered. Thus, it is difficile to distinguish between them. Accordingly, we have ranked the obtained frequent n-grams and itemsets in descending order of their supports. Afterwards, we assign for each item and itemset a weight referring to its importance in the dataset. This weight can be calculated as follow:

$$Weight(Itemset\ I) = \frac{(N+1) - Rank(I)}{(N+1)}$$

Where N is the number of the extracted frequent n-grams or itemsets and Rank is the order of the n-gram or itemset according to its support. For example, the following transaction contains 3 n-grams (1 unigram and 2 bigrams) and 2 itemset having the pursuant supports. On that account, their weights will be as follow (Table 1):

These previous steps are carried out similarly for the dataset appropriated to the non-radical community. So, we obtain likewise for each n-gram and itemset extracted from the non-radical content, their assigned weight. For instance, the following transaction contains 3 n-grams (unigrams) and 3 itemsets extracted from the non-radical context and having the pursuant supports. Hence, their weights will be presented in Table 2:

Table 1. Example of radical itemsets, their supports and their weights

Itemsets and their English translation	Supports	Weight $_{\text{Radical}}$
{باقية تتمدد, قتل } {Residual stretch,kill}	0,4	0,83
{جيش إسلام} {Army of Islam}	0,35	0,66
{قصف، مجاهد} {Shell, mujahid}	0,3	0,5
{جيش عزة, قتل} {Army of Azza, kill}	0,22	0,33
{دولة} {Country}	0,1	0,16

Table 2. Example of non-radical itemsets, their supports and their weights

Itemsets and their English translation	Supports	Weight $_{\text{NonRadical}}$
{خير} {benevolent}	0,5	0,86
{دولة} {Country}	0,3	0,71
{سوريا،سلام} {Syria, peace}	0,25	0,57
{سعادة} {happiness}	0,2	0,43
{مكافحة،إرهاب} {Counter, terrorism}	0,15	0,29
{عالم،حب} {world, love}	0,11	0,14

Hence, we have obtained at this stage two types of n-grams and itemset: (1) a set of n-grams and itemsets designing the radical context and (2) a set of n-grams and itemsets designing the non-radical context.

3.3 Violent Vocabulary Extraction

After the frequent n-grams and itemsets extraction step, we proceed the dangerous vocabulary mining task. At this stage, we aim to explore the obtained frequent n-grams and itemsets of both radical and non-radical contexts in order to extract the violent vocabulary employed by the extremists. Several frequent n-grams and itemsets are used by both radical and non-radical organizations. Thus, we cannot categorize these common n-grams/itemsets as radical or non-radical. For this purpose, we assign for each one a violence degree referring to its degree of danger. The danger degree is calculated as follow:

$$ViolenceDegree(Itemset\ I) = weight(I)_{Radical} - weight(I)_{NonRadical}$$

Where weight(I)$_{\text{Radical}}$ is the weight of the n-grams and itemset I extracted from the radical context and weight(I)$_{\text{NonRadical}}$ is the weight of the n-grams and itemset I in the non-radical context.

For example, following the two previous examples represented in Tables 1 and 2, we obtain the following final vocabulary presented in Table 3 and composing by the set of n-grams and itemsets accompanied by their violence' degrees as well as their annotation:

Table 3. Example of itemsets composing the final vocabulary

Itemsets and their English translation	Weight	Weight	Violence	Class
	Radical	NonRadical	Degree	
{باقية تتمدد, قتل } {Residual stretch,kill}	0,83	0	0,83	Radical
{جيش إسلام} {Army of Islam}	0,66	0	0,66	Radical
{قصف، مجاهد} {Shell, mujahid}	0,5	0	0,5	Radical
{جيش عزة, قتل} {Army of Azza, kill}	0,33	0	0,33	Non-Radical
{دولة} {Country}	0,16	0,71	-0,55	Non-Radical
{خير} {benevolent}	0	0,86	-0,86	Non-Radical
{سوريا، سلام} {Syria, peace}	0	0,57	-0,57	Non-Radical
{سعادة} {happiness}	0	0,43	-0,43	Non-Radical
{مكافحة ،إرهاب} {Counter, terrorism}	0	0,29	-0,29	Non-Radical
{عالم،حب} {world, love}	0	0,14	-0,14	Non-Radical

We can note from the previous table that n-grams and itemsets having positive degree of danger are annotated as radicals. Yet, those which have negative degree are considered as non-radicals. As a result, we obtain the violent vocabulary of radical users on social networks. This vocabulary can be explored later in order to detect extremists' users spreading on social networks.

4 Evaluation

Each data mining methodology should inevitably be evaluated in order to verify its performance in reaching the targeted aim. That's why, we present in this section, the used experiments to evaluate the performance of our methodology to collect the radical vocabulary shared by extremists on social networks. Thus, we have used different library and APIs on the RStudio platform that require a set of development tools to automate their main tasks so that we perform our dangerous vocabulary collection methodology. Furthermore, we have referred to a domain expert in order to evaluate the annotation of the collected dangerous vocabulary. Moreover, in order to validate our expert's annotation, we have resort to a sociologist which plays the role of a second expert for the aim of estimating the inter-annotation agreement between these two experts. So, the rest of the paper will introduce a set of statistics describing our analyzed data as well as our extracted vocabulary and results.

4.1 Implementation Details

In order to perform the required steps and collect violent vocabulary from social networks, we have explored several packages on the RStudio platform. These libraries are described as follow:

Twitter API: In order to target suspicious data from Twitter, we have used the package rtweet on the Rstudio platform. This library provides a simple interface to the Twitter web API. It provides most functionality of the API, with a bias towards API calls that are very useful in data analysis. This limits namely that it doesn't allow extracting historic data since it gives access to only 3200 posts per user [12].

Youtube API: In order to collect data from users' channels on YouTube, we have used the Tuber package on RStudio framework. In fact, the Tuber package [13] provides access to the YouTube API via R. It permits searching for videos having a particular content, getting their statistics and consulting their comments. This package is not only efficient in terms of results but also simple in its manipulation. However, different limitations are encountered using the Youtube API. This later provides only the 50 recent activities of each user. Moreover, for each channel, it returns maximum 50 subscriptions and for each video, 100 comments.

arabicStemR: Since the Arabic text preprocessing raises several issues to be addressed, we have used for this task the arabicStemR package [14] on R Studio framework. This library allows to preprocess Arabic texts for text analysis. It provides several functionalities as removing numbers, cleaning Latin characters, remove punctuation, remove prefixes and remove suffixes.

arules: In order to extract the frequent n-grams and itemsets of radical and non-radical communities on social networks, we have used the arules package [15] for mining association rules and frequent itemsets on R. Studio platform Indeed, this library provides the infrastructure for representing, manipulating and analyzing transaction data and patterns which are the frequent itemsets and the association rules.

4.2 Evaluation and Results

Following the different step described in our methodology, we have collected 8301 n-grams and itemsets annotated as radical and non-radical. Hence, we will report a static that describe our collected data and our extracted vocabulary. The following sectors contain these statistics. The first sector of the Fig. 2 represents the static about the proportion of radical and non-radical analyzed profiles. However, the second sector of the Fig. 2 represents the static about the proportion of radical and non-radical analyzed data. Figure 3 represents the static about the proportion of radical and non-radical extracted vocabulary.

Statistics that will be present in the previous sectors can be explained by the failure of sharing data from radical communities. This is due to their malicious intent in disguising. For this purpose, although the analyzed radical users are more than those who are non-radical, the collected radical data are few comparing with those which are non-radical. In fact, our extracted vocabulary is composing of several n-grams and

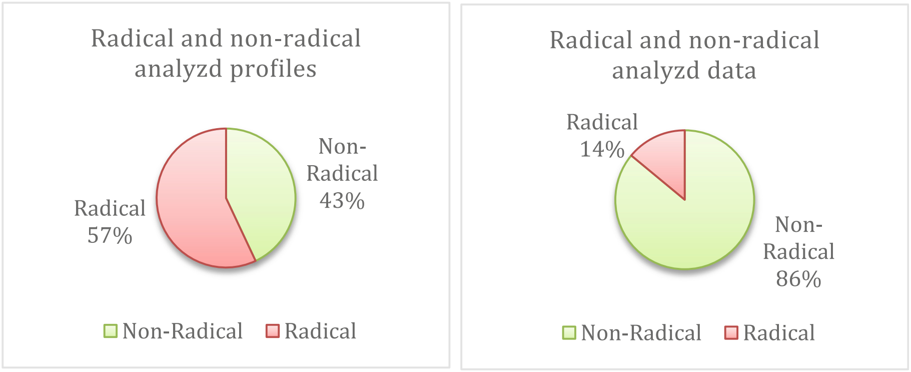

Fig. 2. Statistics about the analyzed profiles and their collected textuel data

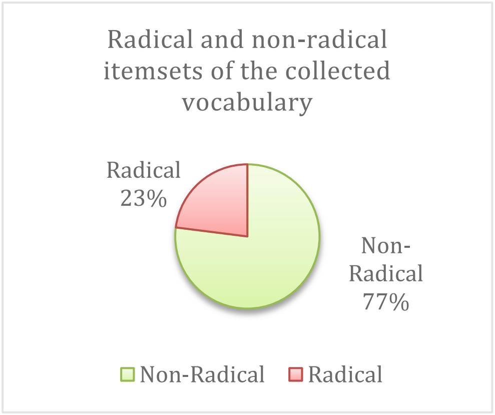

Fig. 3. Statistics about the itemsets composing the extracted vocabulary

itemsets having different sizes. Tables 4 and 5 represent statistics about the obtained n-grams and itemsets size composing the final vocabulary respectively.

Table 4. Statistics about the n-grams composing the extracted vocabulary

n-grams size	Number of n-grams
Unigram	3441
Bigram	47
Trigram	2

In addition, in order to evaluate the effectiveness of the vocabulary annotation provided by our method, we have resort to a domain expert who verifies for each collected n-grams and itemsets whether it is really a radical or non-radical element or

Table 5. Statistics about the itemsets composing the extracted vocabulary

Itemsets size	Number of obtained itemset
9-itemset	1
8-itemset	15
7-itemset	89
6-itemset	285
5-itemset	567
4-itemset	802
3-itemset	1057
2-itemset	1995
1-itemset (n-grams)	3490

not. In fact, the evaluation of our method aims mainly at analyzing its effectiveness and making a judgment that is articulated around a range of criteria. Thus, we can appreciate the quality of our proposed method, according to four of the most important general criteria, which are: recall, precision, F-measure and accuracy. The results given through the expert annotation are represented in Table 6.

Table 6. Results of our proposed methodology

Evaluation measures	Value
Accuracy	0.945
Recall	0.976
Precision	0.951
F-measure	0.96

The obtained results demonstrate the performance of our proposed algorithm, in terms of accuracy, recall, precision and F-measure. Thus, we can confirm that our methodology is highly effective in extracting violent vocabulary shared by radical communities on social networks.

At this stage, we refer to two other sociologists in order to estimate the inter-annotation agreement between them and our expert. This agreement rate is dependent on the number of information having the same annotation by these experts on a test corpus composing of 1000 itemsets composing the extracted vocabulary. To do so, we have explored the Cohen's kappa coefficient (κ) [16]. This later can be calculated as follow:

$$k = \frac{P_0 - P_c}{1 - P_c}$$

Where Po is the proportion of observed agreements and Pc is the proportion of agreements expected by chance.

So, we obtain 0.756 as a Cohen's Kappa coefficient between these two sociologists and our expert. Thus, according to [17], the intra-agreement between these two experts is rather strong. These results show the effectiveness of our methodology in the extraction of violent vocabulary shared by radical communities on social networks. Moreover, our method can be adapted to work for different language since we have explored the n-grams models. This can be considered as an important benefit of our methodology.

5 Conclusion

To crown all, we presented, in this paper, a new methodology for violent vocabulary extraction from social networks. Our methodology is based mainly on collecting a set of extremists and non-extremist users. Then, we focused on their textual content and extracted the frequent n-grams and itemsets appropriated to each community. Finally, we selected the violent vocabulary from the obtained results. The experts' evaluation prompt that our methodology can extract efficiently the radical vocabulary frequently shared by extremists. In the future, we plan to explore our methodology in order to detect dangerous users spreading on social networks. Moreover, we aim to explore our methodology in order to pick up the psyching out of users and extract terrorist communities on social network. In brief, our methodology serves to fight against the violence and the online radicalization.

Acknowledgements. This publication was made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Serrat, O.: Social network analysis. In: Serrat, O. (ed.) Serrat, O. Knowledge solutions, pp. 39–43. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-0983-9_9
2. Kumar, N., Srinathan, K.: Automatic keyphrase extraction from scientific documents using N-gram filtration technique. In: Proceedings of the Eighth ACM Symposium on Document Engineering, pp. 199–208 (2008)
3. Bednár, P.: Vocabulary matching for information extraction language. In: IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 149–152 (2017)
4. Rekik, A., Jamoussi, S.: Deep learning for hot topic extraction from social streams. In: Abraham, A., Haqiq, A., Alimi, A.M., Mezzour, G., Rokbani, N., Muda, A.K. (eds.) HIS 2016. AISC, vol. 552, pp. 186–197. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52941-7_19
5. McCormick, T.H., Lee, H., Cesare, N., Shojaie, A., Spiro, E.S.: Using Twitter for demographic and social science research: tools for data collection and processing. Sociol. Methods Res. **46**(3), 390–421 (2017)
6. Brun, A.: Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole. Ph.D. Nancy 1 (2003)

7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Zhang, Y., Zhang, P., Li, T., Yan, Y.: An unsupervised vocabulary selection technique for Chinese automatic speech recognition. In: Spoken Language Technology Workshop (SLT), pp. 420–425 (2016)
9. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**(4), 537–550 (1994)
10. Maergner, P., Waibel, A., Lane, I.: Unsupervised vocabulary selection for real-time speech recognition of lectures. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4417–4420 (2012)
11. Abid, A., Ameur, H., Mbarek, A., et al.: An extraction and unification methodology for social networks data: an application to public security. In: Proceedings of the 19th International Conference on Information Integration and Web-based Applications and Services, pp. 176–180 (2017)
12. Gentry, J.: Package 'twitteR'. http://cran.r-project.org/web/packages/twitteR/index.html. Accessed 29 Aug 2016
13. Sood, G.: Package 'tuber'. http://cran.r-project.org/web/packages/tuber/index.html. Accessed 28 May 2017
14. Nielsen, R.: Package 'arabicStemmeR'. http://cran.r-project.org/web/packages/arabicStemmeR/index.html. Accessed 7 Feb 2017
15. Hahsler, M., et al.: Package 'arules'. http://cran.r-project.org/web/packages/arules/index.html. Accessed 7 Feb 2018
16. Sim, J., Wright, C.C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. **85**(3), 257–268 (2005)
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics, pp. 159–174 (1977)