Framework to Create Sentiment Lexicons For Extremism Detection in Text Documents

Vijay
Department of Computer Science and Engineering
Bhagwant University
Sikar Road Ajmer, Rajasthan
vijay.movfrose@gmail.com

Pushpneel Verma
Department of Computer Science and Engineering
Bhagwant University
Sikar Road Ajmer, Rajasthan
pushpneelverma@gmail.com

2023 International Conference on IoT, Communication and Automation Technology (ICICAT) | 979-8-3503-0282-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICAT57735.2023.10263761

***Abstract*— Hate speech or extremism detection among the text posts of social media is a major problem nowadays. Hate speech or extremism in text documents can be detected using machine learning methods or using sentiment lexicons. Machine learning methods require already classified data for training. In this paper we have proposed a framework to create sentiment lexicons for any language. In these lexicons terms will be added manually. The sentiment lexicons created by this framework may contain n-grams also where n>= 1. As the proposed framework is used for creating sentiment lexicons for extremism or hate speech detection, it is better to include only negative terms in sentiment lexicons**.

Keywords—Sentiment analysis, Hate speech detection, Sentiment Lexicons.

I. Introduction

The major drawback of machine learning methods for extremism detection is that large amount of already classified data is required for training the model. The performance of machine learning based models depends on the size and accuracy of quality dataset. Lexicon based methods are used for sentiment analysis of text data. Lexicon based methods evaluate the sentiments of text documents by using sentiment lexicons. These methods do not require already classified dataset for training. The sentiment lexicons used by lexicon based methods is either generated automatically or manually created. Adverbs and adjectives are used to evaluate the sentiment of documents. The sentiment lexicon contains those words which express the feelings and opinion of people and it also contains sentiment value or polarity of those words.

Lexicon based methods are considered as rule-based methods. In these methods text documents are tokenized. The stop words and punctuations are removed from text documents and token are formed of words. Lexicon based methods are easy to implement in comparison to machine learning based models. Lexicon based methods can be used detecting extremist text documents.

The lexicon-based methods of sentiment analysis can be classified into two categories: 1. Dictionary based methods and 2. Corpus based methods. Corpus based methods are further classified into two categories i.e., statistical methods and semantic methods [1].

A. Dictionary based methods

In this approach few words are selected initially to create a dictionary. Then this dictionary is expanded by including synonyms and antonyms of these words. For finding synonyms and antonyms, dictionaries like WordNet or SentiWordNet are used.

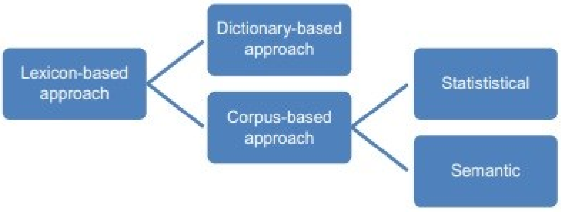


Fig. 1. Types of Lexicon-based methods

B. Corpus-based methods

In these methods the sentiment polarity of a term is found by its semantic distance from seed words. These methods consider syntactic patterns or co-occurrence statistics. In these methods context of a word may be used to assign polarity. These methods can be used to generate domain specific sentiment lexicons. The generated sentiment lexicon is not limited to the formal dictionary terms. It may contain informal terms and social media slangs [2]. There are two approaches in corpus-based methods:

- Semantic approach: In this approach the words which are semantically nearer are assigned alike sentiment values. Semantically nearer words can be found by getting synonyms and antonyms of a word [1].
- Statistical approach: In this approach the polarity of a word is determined as positive if it is found commonly in a text document with positive polarity. The polarity of a word is determined as negative if it is found frequently in text document with negative polarity. The polarity of a word is considered neutral if it has same frequency in both positive and negative text documents.

On social media like Facebook, it has been found that people usually write in Hindi language by using English script. We developed a framework to create sentiment lexicon manually which contains sentiment terms of Hindi language written by using English script. Our purpose is to detect only extremist text posts on social media. Therefore, we need to include only negative sentiment terms in the lexicon. Detecting positive sentiment terms in text documents does not have role in the detection of extremism in text documents. It is useless to include positive sentiment terms in lexicons for extremism detection. There is no standard way to write Hindi terms by using English script. Hence, negative sentiment terms of English language cannot be translated automatically for such lexicons. Moreover, there are many slangs in Hindi language which cannot be automatically translated into English language because if translate them word by word into English language then we will get a text in English with incorrect meaning. The proposed framework is not only for creating sentiment lexicons for Hindi language but it can also be used for creating sentiment lexicons for other languages. Sentiment lexicons created by this framework will contain not only unigram but it can also contain n-grams, where n>1.

II. Related Work

Al-Ayyoub [3] built a sentiment lexicon for the sentiment analysis of Arabic texts. This lexicon contained 120,000 Arabic words. To construct this lexicon, they took some Arabic stems from dataset of Abuaiadh [4]. Then they collected lots of article from different Arabic news websites using crawler. From these articles Arabic stem were extracted by using a tool which was also built by them. They translated these Arabic stems into English and then by using sentiment lexicons of English language they found sentiment value of each Arabic word of their lexicon. They built a tool for sentiment analysis relying on predicate calculus. The performance of their proposed lexicon-based sentiment analysis tool was evaluated by them by performing experiments. They compared their proposed method with keyword-based method of sentiment analysis and found that their proposed method delivered better prediction accuracy.

Taj et al. [5] performed sentiment analysis using WordNet dictionary. For their experiments they used a BBC news dataset which contains 2225 documents. News articles of the dataset belong to five domains: tech, entertainment, sport, business and politics. In their experiments they had found that articles belonging to sports and business domains were more positive and articles belonging to tech and entertainment domains were more negative.

Dey et al. [6] created a n-gram sentiment lexicon named Senti-N-Gram. They compared the performance of their proposed method with a method using sentiment dictionary VADER and SO-CAL method which is used to find sentiment scores of n-grams. The performance of Senti-N-Gram was found better than these two methods. For evaluating performances, they performed experiments on two datasets namely: Movie review [7] and Epinion [8].

Asghar et al. [9] created a sentiment lexicon for the Urdu language. They used Bing Liu's list containing more than 6,000 sentiment words, Oxford English to Urdu dictionary, SentiWordNet and Urdu sentiment lexicon Chaoticity. The average sentiment scores of sentiment words from the list of sentiment words were found using SentiWordNet. Using a bilingual dictionary these sentiment words were translated from English to Urdu. Then average sentiment scores which had been evaluated for English sentiment words were assigned to their translated equivalent Urdu sentiment words. They got 1002 negative Urdu sentiment words and 1044 positive Urdu sentiment words by using this method. They also acquired 4728 negative sentiment words and 2607 positive sentiment words of Urdu from another list created by Awais. The lexicon proposed by them contained 3,651 Urdu words labeled as positive and 5730 Urdu words labeled as negative. They also collected Urdu modifiers and assigned proper sentiment scores to them.

Sajeetha et al. [10] proposed a method for the expansion of sentiment lexicon. They used fast Text and Word2Vec word embeddings. Initially they took a seed list for the expansion of sentiment lexicon. This list contained 5598 negative words and 2951 positive words. First, they used Word2Vec to acquire relevant words, then they used fastText word to find lexically related words. They created a lexicon named UJ_Lex_Poscontaining 10537 positive words and UJ_Lex_Neg containing 12664 negative words. Then they performed sentiment analysis of Tamil texts using these lexicons along with the list of negational words and conjunctions. They evaluated their proposed method on UJ_MovieReviews dataset and got accuracy of 88 +-0.14%.

Mataoui et al. [11] created three sentiment lexicons for Algerian Arabic language and a dataset which were annotated manually. Palanisamy et al. [12] created a lexicon for SemEval-2013 Task 2. Then they classified tweets by using that lexicon as negative or positive. On test dataset they got 0.8004 value of F-score. Lexicon was created by them manually.

Keshavaraz and Abadeh [13] proposed a genetic algorithm for creating adaptive sentiment lexicons for text classification. The sentiment lexicon was created by their proposed method for training dataset and then this lexicon was used on test data. Sentiment scores were assigned to the sentiment words present in lexicon. Words appearing more than two times in a tweet of training dataset were considered as sentiment word. Additional features were also used along with meta –level feature extracted by proposed algorithm. They performed experiments on six datasets to evaluate proposed method and got significantly good results.

Sabra et al. [14] proposed a method to create an Arabic sentiment lexicon. Initially they created a wordlist of 9 negative sentiment words and 9 positive sentiment words. They expanded the list by using English WordNet with synset relations. The seed list was expanded twice, once for negative sentiment words and another for positive sentiment words. They also used SAMA (Standard Arabic Morphological Analyzer) databases. The sentiment lexicon created by them contained around 75,000 sentiments words with sentiment value negative, positive or neutral. They evaluated the performance of lexicon by performing a classification experiment on a dataset of movie reviews in Arabic. This dataset contained movie reviews which were divided into two categories: positive and negative. In their experiment they got 65% of average F measure.

Oliveira at al. [15] proposed a method for automatic creation of sentiment lexicon. For creating sentiment lexicon, they used labeled messages from social media website StockTwits and statistical measures: Information Gain, Weighted Class Probability, Class Percentage and TF-IDF. They created a sentiment lexicon for the stock market domain. They compared the performance of created lexicon with other six lexicons i.e., Harvard General Inquirer, Opinion Lexicon, Macquarie Semantic Orientation Lexicon, MPQA Subjectivity Lexicon, SentiWordNet and Financial Sentiment Dictionaries. The performance of the created lexicon was found competitive.

Abdulmohsen et al. [16] created a sentiment lexicon named SauDiSenti for the Saudi dialect. This lexicon contained words, phrases and Saudi dialects. They also created a dataset containing 1500 tweets divided into three categories: negative, positive and neutral. They compared the performance of SauDiSenti with AraSenTi. For comparing performances, they considered F measure, precision and recall. They found that performance of AraSenTi was better than SauDiSenti when neutral tweets were not considered. When all negative, positive and neutral tweets were considered, the performance of SauDiSenti was found better than AraSenTi.

Anastasia and Evgeny [17] suggested that sentiment lexicons containing collocations can perform better than the sentiment lexicons containing individual words. They argued that some collocations' meaning is not related to the meaning of words they consist of. They created sentiment lexicons containing collocations by using a semi-automatic approach. They created sentiment lexicons for these domains: reviews of books, cars,movies, computers, banks, phones, house appliances, music, hotels and restaurants. These sentiment lexicons were unioned to form a single universal lexicon called SentiRusColl. SentiRusColl contained 6,350 sentiment collocations in which 2508 were negative and 3842 were positive. They revealed that union of SentiRusColl and sentiment lexicon RuSentiLex delivered better performance for sentiment analysis.

Asghar et al. [18] created a domain specific sentiment lexicon for sentiment analysis of health reviews regarding a drug or medication. The method they used for creation of the lexicon was a combination of boot-strapping and corpus-based techniques. They created a seed list of sentiment terms and expanded it by a boot-strapping method. They removed irrelevant through co-reference PMI measure. They used UMLS (Unified Modeling Language System) tags for labeling terms as non-medical terms or medical terms. Sentiment classes were assigned to sentiment words using Class-based probability. For assigning sentiment scores to sentiment words, they proposed an enhanced term weighting scheme. They also demonstrated the efficacy of proposed SentiHealth.

Dehkharghani et al. [19] created a sentiment lexicon for Turkish language named as SentiTurkNet. For creating SentiTurkNet they used English WordNet, SentiWordNet and SenticNet. They also used Turkish WordNet which contains 15,000 synsets.

Yekrangi and Abdolvand [20] created a domain specific sentiment lexicon for the sentiment analysis of financial markets. To create this lexicon, they used hybrid approach including both dictionaries based and corpus-based methods. To create this lexicon, they used wide related datasets and dictionaries. They found seed wordlist using a dataset containing new articles of Reuters and Bloomberg. They selected a set of words on the basis of statistical feature and then verified these words whether they were related to financial context or not and removed unrelated words. For verification they used Financial Times and Loughran-McDonald dictionaries. They evaluated PMI (point wise mutual information) index for each verified word. The sentiment score of each word was evaluated using its PMI. After getting sentiment scores words were categorized into positive and negative classes. Then they expanded the wordlist using Oxford and WordNet dictionaries. SentiWordNet and InvestorWords dictionaries were used for the verification of wordlist. The experiments performed to evaluate the performance of created lexicon demonstrated a high correlation between sentiments obtained using created lexicon and market trends.

Neelam et al. [21] performed sentiment analysis of blogs of different domains written in Urdu using supervised machine learning based methods and lexicon-based method. They used three supervised machine learning based classifiers i.e., K Nearest Neighbor, Support Vector Machine and Decision Tree. In lexicon-based technique for sentiment analysis they created an Urdu sentiment lexicon. For evaluating the performances of supervised machine learning based classifiers and lexicon-based method they used a dataset containing 6025 sentences taken from 151 blogs of 14 different categories. The 14 different categories of blogs along with the number of blogs belonging to each category were: Language and Literature (6), Current Affairs (32), Health (10), Ethics (10), History (8), General Knowledge (13), Psychology (10), Personalities (3), Economy (6), Religion (16), Technology (10), Sports (5), Humor and Satire (5) and Politics (17). The results of their experiments demonstrated that the performance of lexicon-based method was better than supervised machine learning based classifiers in terms of time, efforts, recall, accuracy, precision and F-measure.

Gavilanes et al. [22] created a sentiment lexicon for emojis. For creating this lexicon, they considered both the definitions of emojis given in Emojipedia and the texts written with emojis, and then performed sentiment analysis of both.

Machado et al. [23] introduced a context sensitive sentiment lexicon in Portuguese and also evaluated. The lexicon they built was Lexicon ReLi (LexReLi) and it was specialized in determining the polarity of book reviews taken from ‘skoob’ website. They created a dataset by using ReLi corpus. The ReLi contains 1600 Portuguese book reviews of fourteen books. They extended the ReLi corpus to form the dataset. The created extended dataset contained 6698 reviews, 980640 words and 511478 phrases. The preprocessed both ReLi corpus and the corpus they created by extending ReLi. To construct LexRexLi first they selected those sentences in which nouns were close to adjectives. They found the polarity of selected sentences. If the frequency of occurrence of an adjective in positive phrases was more than its frequency of occurrence in negative phrases then the positive polarity was assigned to that adjective. Negative polarity was assigned to an adjective if its frequency of occurrence in negative phrases was more than frequency of occurrence in positive phrases. An adjective was not included in lexicon if the difference between the frequencies of its occurrence in positive sentences and in negative sentences was less than two. They also formed lexicons by combining SentiLex, Opinion Lexicon and Brazilian Portuguese LIWC (Linguistic Inquiry and Word Count) with different combination approaches. To evaluate these lexicons, they performed experiments with the corpus ReLi.

Wijayanti and Arisal [24] built an Indonesian sentiment lexicon called SentIL. To create SentIL, first they created a seed lexicon containing sentiment words with significant sentiment polarity. They mapped SentiWordNet with WordNet Bahasa with automatic and hand-checked translation. Only those sentiment words whose sentiment scores were higher than 0.5 were considered by them or creating seed lexicon. If an Indonesian sentiment word mapped to more than one word in English SentiWordNet then the sentiment scores of all these words were averaged and the result was assigned as the sentiment score of Indonesian word. They got a 1959 synset and used it as seed words. To expand seed lexicon, they used dictionary-based techniques and corpus-based techniques for formal words and informal words respectively. In dictionary-based approach they used an Indonesian online dictionary Kateglo to add synonyms and antonyms. In corpus –based technique they used Word2Vec with continuous skip gram and CBOW (continuous bag-of-words) model to find similar words for the expansion of lexicon. In lexicon some emoticons and slang words were also added. They also tuned the sentiment values except of emoticons. The performed experiment to evaluate the performance of created sentiment lexicon SentIL. The result of experiment demonstrated that F1 score and accuracy could be improved for Twitter data and online reviews by SentIL

Yatim et al. [25] proposed a corpus-based method for creating a contextual sentiment lexicon in which as corpora news articles were used. They created a lexicon for Indonesian politic context. The proposed method includes three steps for creating the lexicon. The first step is crawling. They performed crawling on viva.co.id, detik.com, okezone.com, tribunnews.com and kompas.com for collecting corpora. The second step is analysis of term Document Frequency (DF). In this step first they removed stopwords from corpus and then remaining words with their DF scores were stored. Six positive sentiment terms and six negative sentiment terms were chosen from 300 terms of highest DF score by the experts of political domain. These chosen words formed a seed wordlist. The third step is Lexicon building. To build the lexicon they introduced a method called LEXBILD system.

WKWSCI is a sentiment lexicon which was constructed by fourth- and third-year students of the programme of Bachelor of Communication Studies of Nanyang Technological University of Singapore at Wee Kim Wee School of Communication and Information [26]. This sentiment lexicon was coded manually by these undergraduate students. There are 29718 words in this sentiment lexicon in which 3121 sentiment words have positive sentiment polarity and 7100 sentiment words have negative sentiment polarity. The sentiment polarities were assigned to the sentiment words in two phases. In the first phase sentiment words were categorized as positive, negative or neutral. In the second phase positive sentiment words were further sub-classified as very positive, positive and slightly positive. The sentiment values of 3, 2 and1 were assigned to very positive, positive and slightly positive sentiment words. Similarly, negative sentiment words were classified as very negative, negative and slightly negative and assigned them sentiment values of -3, -2 and -1 respectively.

SentiWordNet is a lexical resource which is used for opinion mining and sentiment classification [28]. Esuli and sebastiani created SentiWordNet 1.0 [27]. For research purposes, SentiWordNet 1.0 is publicly available. It was created by annotating WORDNET's synsets automatically. Three numerical scores were assigned to each synset x, i.e., Obj(x), Pos(x) and Neg(x). The numerical scores Obj(x), Pos(x) and Neg(x) specify the degree of objectivity (or neutrality), positivity and negativity respectively of sentiment words of the synset. The sum of all three numeric scores for each synset was 1. Each of these numeric scores has a value between 0.0 and 1.0. SentiWordNet 3.0 is an enhanced version of SentiWordNet 1.0. SentiWordNet 3.0 was created in two steps: Semi-supervised learning and random walk. In semi supervised learning step, first they took two seed sets. One seed set contained all synsets having 7 positive sentiment terms and another seed set contained all synsets having 7 negative sentiment terms. These seed sets were expanded automatically by using WORDNET binary relations. They trained a ternary classifier for the classification of synsets. The ternary classifier classified all WORDNET synsets. Using different radius and learning technologies 8 different ternary classifiers were generated. The final numeric scores assigned to a synset were the average of numeric scores across all eight classifiers. In second step WORDNET 3.0 was considered as a graph and "random walk" which is an iterative process was run over it. In each iteration the "random walk" changed the numeric score values of synsets.

General Inquirer was developed in 1961 at Harvard. It is an IBM 7090 program system [29]. It contains 11789-word senses which are clustered into 182 categories [26].

Hiu and Liu opinion lexicon [30] contained 6790 terms (4783 negative terms and 2006 positive terms). It was built automatically by the machine learning approach on the basis of collected reviews of different domains given by customers in several years

MPQA subjectivity lexicon contained 8222 terms in which 2719 terms were categorized as positive, 591 terms as neutral and 4914 terms as negative [26]. This lexicon contains nouns, adjectives, verbs, adverbs and 'anypos' (refers to any part of speech). This lexicon was created by using a number of sources, some sources were developed manually and some sources were constructed automatically. Riloff and Wiebe [31] added most of the terms in this lexicon.

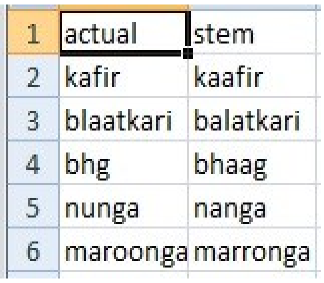
EmoLex (NRC Word-Emotion Association Lexicon) contains sentiment terms of the English language and the association of each sentiment term with eight types of emotions i.e., disgust, trust, anger, joy, fear, sadness, anticipation and surprise and two kinds of sentiments i.e., positive sentiment and negative sentiment are specified. Over 8000 sentiment terms in this lexicon are taken from General Inquirer and from WordNet Affect Lexicon [32] 640 terms are taken. Emolex also contains 200 most frequent unigrams and 200 most frequent bigrams for all part-of-speeches (i.e., verbs, adverbs, nouns and adjectives) of Macquarie Thesaurus [33].

III. METHODOLGY

To create lexicon through proposed framework, one needs to create only two files. One of them is stemming list containing only two columns labeled as actual and stem as shown in Fig. 2.

Stemming is necessary because there is no standard way to write terms of Hindi language in English. People may write a same word with different spellings in English. Therefore, stemming needs to be done before the sentiment analysis of a text document. In stemming the terms present in column labeled as actual are stemmed to their corresponding terms of stem column. For example, according to the stemming list given in Fig. 1 in text "*tujhe nunga kar ke maroonga"* the term "*nunga*" will be stemmed to "*nanga*" and "*maroonga*" will be stemmed to "*marronga*." Second file which is required to create sentiment lexicon is the collection of sentiment terms separated in different columns according to their categories. For example, as shown, in Fig. 2 the sentiment terms are separated into three columns according to their categories i.e., hate, rejection and disgrace. This file contains not only unigrams but it can also contain n-grams where n>1. For example, there is a trigram "akal ke andhe" in disgrace category.

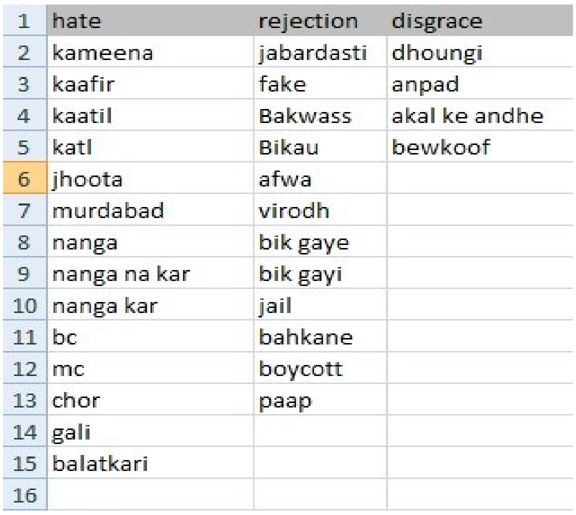
The lexicons created by our proposed framework are used for sentiment analysis of text documents. A text document is first divided into individual words and stemming is performed on these individual words by using the already created stemming list. After stemming individual words of a text document are combined to form the string of words separated by a space. Now, the file containing sentiment terms (including ngrams also where n>1) is used to identify the sentiment terms in a text. Finally, a matrix containing a row for each text document and a column for each sentiment category is returned. The entries in the matrix specify the number of sentiment terms (including ngrams also where n>1) of a particular category available in a text document.



1	actual	stem
2	kafir	kaafir
3	blaatkari	balatkari
4	bhg	bhaag
5	nunga	nanga
6	maroonga	marronga

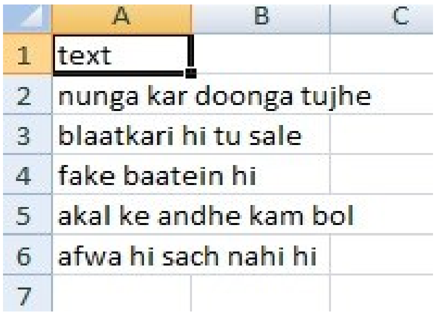
Fig. 2. Stemming list to create sentiment lexicon

Let us consider that a sentiment lexicon for extremism detection is created by using stemming list shown in Fig. 2 and collection of sentiment terms shown in Fig. 3. Now, we have to perform sentiment analysis of text documents shown in Fig. 4 by using this created lexicon.



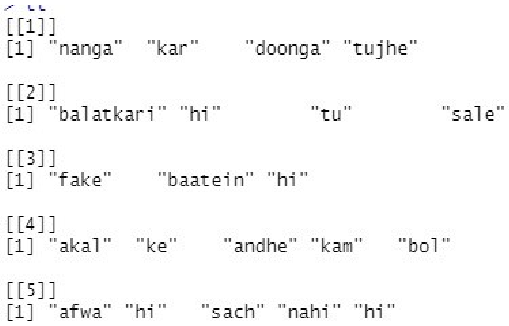
1	hate	rejection	disgrace
2	kameena	jabardasti	dhoungi
3	kaafir	fake	anpad
4	kaatil	Bakwass	akal ke andhe
5	katl	Bikau	bewkoof
6	jhoota	afwa	
7	murdabad	virodh	
8	nanga	bik gaye	
9	nanga na kar	bik gayi	
10	nanga kar	jail	
11	bc	bahkane	
12	mc	boycott	
13	chor	paap	
14	gali		
15	balatkari		
16			

Fig. 3. Collection of sentiment terms



	A	B	C
1	text		
2	nunga kar doonga tujhe		
3	blaatkari hi tu sale		
4	fake baatein hi		
5	akal ke andhe kam bol		
6	afwa hi sach nahi hi		
7			

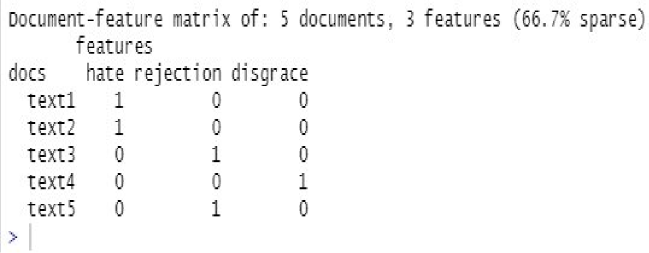
Fig. 4. Text documents



```
[[1]]
[1] "nanga"   "kar"     "doonga" "tujhe"

[[2]]
[1] "balatkari" "hi"         "tu"         "sale"

[[3]]
[1] "fake"    "baatein" "hi"

[[4]]
[1] "akal"   "ke"     "andhe" "kam"    "bol"

[[5]]
[1] "afwa" "hi"    "sach" "nahi"  "hi"
```

Fig. 5. Stemming on given text documents

As shown in Fig. 5 the term "nunga" of first text document has been stemmed to "nanga" and the term "blaatkari" has been stemmed to "balatkari". These individual terms after stemming are joined to form texts again. Finally, by using collection of sentiment terms shown in Fig. 3 a matrix is formed (as shown in Fig. 6) showing number of sentiment terms of a particular category available in a text document. For example, as shown in Fig. 6 the first text document contains only one sentiment term of hate category. The fourth document contains a 3-gram "akal ke andhe" of disgrace category. Therefore, in fourth row of matrix the column labeled as disgrace contains the value 1.



```
Document-feature matrix of: 5 documents, 3 features (66.7% sparse)
       features
docs    hate rejection disgrace
  text1    1         0        0
  text2    1         0        0
  text3    0         1        0
  text4    0         0        1
  text5    0         1        0
> |
```

Fig. 6. Matrix formed after sentiment analysis of text documents

IV. Conclusion

The proposed framework can create sentiment lexicons for any language written in English script. Lexicon based methods for hate speech detection does not require training dataset. Stemming is also performed by the sentiment lexicons created by our proposed framework.

References

[1] Gupta, N., & Agrawal, R. (2020). Application and techniques of opinion mining. Hybrid Computational Intelligence, 1–23. doi:10.1016/b978-0-12-818699-2.00001-9

[2] Darwich, Mohammad & Mohd Noah, Shahrul Azman & Omar, Nazlia & Osman, Nurul. (2019). Corpus-Based Techniques for

Sentiment Lexicon Generation: A Review. Journal of Digital Information Management. 17. 296. 10.6025/jdim/2019/17/5/296-305.

[3] Al-Ayyoub, Mahmoud, Safa Bani Essa, and Izzat Alsmadi. "Lexicon-based sentiment analysis of Arabic tweets." *Int. J. Soc. Netw. Min.* 2.2 (2015): 101-114.

[4] Abuaiadh, D. (2011) Dataset for Arabic Document Classification [online] http://diab.edublogs.org/dataset-for-arabic-document-classification/ (accessed June 2013).

[5] Taj, Soonh; Shaikh, Baby Bakhtawer; Fatemah Meghji, Areej (2019). [IEEE 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) - Sukkur, Pakistan (2019.1.30-2019.1.31)] 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) - Sentiment Analysis of News Articles: A Lexicon based Approach. , (), 1–5. doi:10.1109/ICOMET.2019.8673428

[6] Dey, Atanu; Jenamani, Mamata; Thakkar, Jitesh J. (2018). Senti-N-Gram: An n-gram lexicon for sentiment analysis. Expert Systems with Applications, (), S095741741830143X–. doi:10.1016/j.eswa.2018.03.004

[7] Pang, B. and Lee, L. (2004), A sentimental education: Sentiment analysis using subjectivity, in 'Proceedings of ACL', pp. 271–278.

[8] Taboada, M. and Grieve, J. (2004), Analyzing appraisal automatically, in 'Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press'.

[9] Asghar, Muhammad Zubair; Sattar, Anum; Khan, Aurangzeb; Ali, Amjad; Masud Kundi, Fazal; Ahmad, Shakeel (2019). Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. Expert Systems, (), e12397–. doi:10.1111/exsy.12397

[10] S. Thavareesan and S. Mahesan, "Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts," *2020 Moratuwa Engineering Research Conference (MERCon)*, 2020, pp. 272-276, doi: 10.1109/MERCon50084.2020.9185369.

[11] Mataoui MH, Zelmati O, Boumechache M. A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic. Research in Computing Science. 2016 Dec;110(1):55-70.

[12] Palanisamy, Prabu, Vineet Yadav, and Harsha Elchuri. "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis." *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 2013.

[13] Srivastav, S., Singh, P.K. & Yadav, D. A novel approach to solve exact matching problem using multi-splitting of text patterns. Int J Syst Assur Eng Manag 14, 1457–1466 (2023). https://doi.org/10.1007/s13198-023-01948-7

[14] Sabra, K. S., Zantout, R. N., Abed, M. A. E., & Hamandi, L. (2017). Sentiment analysis: Arabic sentiment lexicons. 2017 Sensors Networks Smart and Emerging Technologies (SENSET). doi:10.1109/senset.2017.8125054

[15] Oliveira, Nuno; Cortez, Paulo; Areal, Nelson (2014). [ACM Press the 18th International Database Engineering & Applications Symposium - Porto, Portugal (2014.07.07-2014.07.09)] Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS '14 - Automatic creation of stock market lexicons for sentiment analysis using StockTwits data. , (), 115–123. doi:10.1145/2628194.2628235

[16] Al-Thubaity A, Alqahtani Q, Aljandal A. Sentiment lexicon for sentiment analysis of Saudi dialect tweets. Procedia computer science. 2018 Jan 1;142:301-7.

[17] Kotelnikova A, Kotelnikov E. SentiRusColl: Russian collocation lexicon for sentiment analysis. InConference on Artificial Intelligence and Natural Language 2019 Nov 20 (pp. 18-32). Springer, Cham.

[18] Asghar, M.Z., Ahmad, S., Qasim, M. *et al.* SentiHealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus* 5, 1139 (2016). https://doi.org/10.1186/s40064-016-2809-x

[19] Dehkharghani, R., Saygin, Y., Yanikoglu, B. *et al.* SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Lang Resources & Evaluation* **50,** 667–685 (2016). https://doi.org/10.1007/s10579-015-9307-6

[20] Yekrangi, M., & Abdolvand, N. (2020). Financial markets sentiment analysis: developing a specialized Lexicon. Journal of Intelligent Information Systems. doi:10.1007/s10844-020-00630-9

[21] Neelam Mukhtar, Mohammad Abid Khan, Nadia Chiragh, Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains,Telematics and Informatics, Volume 35, Issue 8, 2018, Pages 2173-2183, ISSN 0736-5853, https://doi.org/10.1016/j.tele.2018.08.003.

[22] Fernández-Gavilanes M, Juncal-Martínez J, García-Méndez S, Costa-Montenegro E, González-Castaño FJ. Creating emoji lexica from unsupervised sentiment analysis of their descriptions. Expert Systems with Applications. 2018 Aug 1;103:74-91.

[23] Machado MT, Pardo TA, Ruiz EE. Creating a Portuguese context sensitive lexicon for sentiment analysis. InInternational Conference on Computational Processing of the Portuguese Language 2018 Sep 24 (pp. 335-344). Springer, Cham.

[24] Wijayanti R, Arisal A. Automatic Indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). 2021 Mar 2;20(1):1-6.

[25] M. A. F. Yatim, Y. Wardhana, A. Kamal, A. A. R. Soroinda, F. Rachim and M. I. Wonggo, "A corpus-based lexicon building in Indonesian political context through Indonesian online news media," *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 347-352, doi: 10.1109/ICACSIS.2016.7872794.

[26] S. Srivastav, P. K. Singh, and D. Yadav, "A Method To Improve Exact Matching Results in Compressed Text Using Parallel Wavelet Tree," Scalable Comput., vol. 22, no. 4, pp. 387–400, Nov. 2021.

[27] Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pages 417–422, Genova, IT.

[28] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[29] Stone PJ, Hunt EB. A computer approach to content analysis: studies using the general inquirer system. InProceedings of the May 21-23, 1963, spring joint computer conference 1963 May 21 (pp. 241-256).

[30] Hu M and Liu B. Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, Seattle, WA, 22–25 August 2004, pp. 168–177. New York: ACM

[31] Riloff E and Wiebe J. Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on empirical methods in natural language processing, Sapporo, Japan, 11–12 July 2003, pp. 105–112. Stroudsburg, PA: Association for Computational Linguistics.

[32] Strapparava C and Valitutti A. Wordnet-Affect: An affective extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation LREC-2004, Lisbon, Portugal, 2004, pp. 1083–1086

[33] Bernard J (ed.). The Macquarie Thesaurus. Sydney: Macquarie Library, 1986.