

Text pre-processing of multilingual for sentiment analysis based on social network data

Neha Garg, Kamlesh Sharma

Department of Computer Science and Engineering, Manav Rachna International Institute of Research and Studies, Faridabad, India

Article Info

Article history:

Received Mar 25, 2021

Revised Jul 15, 2021

Accepted Aug 1, 2022

Keywords:

Code-switch

Linguistic-switching

Machine learning

Multilingual

Pre-processing

Sentiment analysis

ABSTRACT

Sentiment analysis (SA) is an enduring area for research especially in the field of text analysis. Text pre-processing is an important aspect to perform SA accurately. This paper presents a text processing model for SA, using natural language processing techniques for twitter data. The basic phases for machine learning are text collection, text cleaning, pre-processing, feature extractions in a text and then categorize the data according to the SA techniques. Keeping the focus on twitter data, the data is extracted in domain specific manner. In data cleaning phase, noisy data, missing data, punctuation, tags and emoticons have been considered. For pre-processing, tokenization is performed which is followed by stop word removal (SWR). The proposed article provides an insight of the techniques, that are used for text pre-processing, the impact of their presence on the dataset. The accuracy of classification techniques has been improved after applying text pre-processing and dimensionality has been reduced. The proposed corpus can be utilized in the area of market analysis, customer behaviour, polling analysis, and brand monitoring. The text pre-processing process can serve as the baseline to apply predictive analysis, machine learning and deep learning algorithms which can be extended according to problem definition.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Neha Garg

Department of Computer Science and Engineering, Manav Rachna International Institute of Research and Studies

Sector-43, Delhi Surajkund Road, Faridabad, Haryana, 121005, India

Email: gargsneha99@gmail.com

1. INTRODUCTION

Sentiment analysis (SA) is a field of natural language processing (NLP) for analyzing the opinion, expression and attitude of the user towards an entity which can be an individual, a place, an event, product, and issue or a discussion. The markets, firms and organizations utilize sentiment analysis to attract and satisfy more customers. Political parties of different countries use the SA data to achieve the satisfaction of the citizens or to know the feedback of its people about its administration and policies. SA offers more challenging opportunities to develop new applications where sentiments of customers, viewers or users play a vital role. The high level of precision of any decision can bring an organization at the top position among its competitors and can make any organization fall from the roof [1] with the advanced technology and increasing the use of gadgets, sentiments are not only obtained from orthodox feedbacks only but may also be in the form of audios, videos, images, texts, micro-blogs, tweets, posts and comments on social media or emoticons.

Analyzing such heterogeneous data in real-time with appreciable level of precision has always been a challenge. Most of the organizations use some third-party tool such as WorkForce, Glassdoor or sometimes make a separate cell to analyze this data which provide them genuine sentiment analysis followed by an accurate decision making. Inaccurate SA or poor level of accuracy in SA may prove to be disastrous for any

organization in present time of cut throat market competition. A lot of improvement in SA and its applications has been made in recent past.

2. LITERATURE REVIEW

More often the information available on social media sites is not well formed in structure like it has abbreviations, miss-spelling, emoticons, slang language, and html tag which turn into a difficult task for sentiment analysis [2], [3]. This type of data increases the dimensionality which may leads to bad performance of sentiment analysis (SA) techniques [4]. To reduce the dimensionality, before applying SA technique, the data has been passed through two phases i.e., data cleaning and pre-processing. The accuracy of SA techniques majorly depends on these phases.

Many techniques have been used for data classification and data clustering. It is found that in most of the cases, hybrid methods are used for classification and clustering based on the problem at hand [5]. Classification techniques are the sub-category of supervised techniques where some kind of labelled data is used to train the dataset based on which outcome is provided. The degree of accuracy is highly dependent on the accuracy of training data [6]. A semi-supervised technique, regularized least squares (RLS), was introduced to represent unlabeled and labelled data on bipartite graph representation to analyze the sentiment of documents as well as of words [7]. The blogs considering enterprise software products, politics and movie reviews were considered and the technique produced 90% accuracy. Using supervised techniques such as support vector machine (SVM), multinomial Naïve Bayes (MNB) and maximum entropy (Max Ent), Erik and Francine [8] worked on multilingual unformatted dataset of English, Dutch and French languages and achieved 58% accuracy. In this technique, the implementation of automated analysis provided impending benefits such as word-of-mouth marketing, real-time response and neutral fetching of information. The smallness of dataset did not help much in managing noise. Term-document matrix was employed for tri-factorization by making some simple updates in rules where sentiment lexicons were used as the first set of constraints. Second set of constraints retained domain specific supervisions [9].

Blogs having four different dimensions of discussion i.e., software product, politics, amazon product reviews and movie reviews were worked upon. In a new approach rule base classification and machine learning approaches were coupled together and 50% accuracy was achieved [2]. A compact semi-supervised classifier was introduced in which classifiers were assigned according to the type of text. In this pipeline approach, 10-fold cross validation was performed which resulted in higher efficiency but consumed more time.

Li *et al.* [10] had utilized SVM and NB techniques for reviews of Books, DVD, kitchen appliances and electronic items. The data was split into two categories, personal and impersonal text as co-training data. This, along with improving baseline accuracy [11], reduced classification noises and needed no proper syntactical rules. Lack of labelled data was treated well by dividing imbalance population into multiple sets of balanced population for sentiment classification and multiple iteration improved performance [12]. A new approach was provided for active-learning in multi-domain framework. The term frequency method was used to weigh the features along with LIBLINEAR SVM. This method was compared with spam mail filtering, newsgroup classification and sentiment classification where human efforts were reduced by 33.2%, 42.9% and 68.7% respectively.

To classify sentiments of micro blogs, a method was proposed in which machine learning was combined with domain specific techniques and a system called opinion miner was introduced [13]. The precision of opinion miner stood at 96%. Lack of stop-criteria to control iteration created unnecessary data sampling. Numerical matrix representation was used for movie reviews and positive or negative reviews were obtained with 89.5% accuracy with NB. SVM enhanced accuracy to 94% [1]. Sentiment analysis was performed on reviews [14]. The performance of unigram with stop word settled at 82.9% and that of without stop words came 83% with positive class. The same was higher for negative class.

Based on deep learning parameters, a model was proposed to address implicit and explicit sentiment factors on text data and used word embedded representation in Vietnamese and English language [15]. The proposed model proved to be better than traditional machine learning methods and provided results up to 87% of sentiment analysis in all available corpora.

Clustering techniques are used to cluster the data based on different parameters and to form groups as per the requirements for business analytics for better decision making [16]. A techniques using k-means clustering was proposed to cluster the document in combination with scoring technique [17]. The movie review dataset was examined with 77.17% accuracy. Spam filtering technique was developed which was based on the vector space model by using text clustering k-means and balanced iterative reducing and clustering using hierarchies (BIRCH) technique [18]. K-means clustering provided better results for smaller data. k nearest neighbours (k-NN) and BIRCH were shown to be good for larger datasets.

Venkatasubramanian *et al.* [19] proposed that without employing syntactic processing, stop words could be used for classification. A semantic clustering algorithm, latent dirichlet allocation (LDA), was

applied and 66% for the polarity of reviews was observed. This method proved to be complementary for the syntactic approach for sentiment analysis. For product reviews, a semi-supervised technique was given where words and phrases belonging to similar domain were grouped under same feature set [20]. The expectation maximization (EM) algorithm based on naïve Bayes was applied on five datasets and results were found to be superior to the 13 baselines which presented current state of art solution. The authors have presented a contextual multimodal method to analyze visual, textual and audio cues of approx 800 utterances for Persian language and achieve a performance of around 91% [21].

An approach was suggested which was based on clustering with term frequency – inverse document frequency (TF-IDF) weighing technique, voting mechanism and important term score [22]. This approach was shown to be efficient, automated, accurate and faster than other supervised learning techniques. Further modifications were introduced in clustering technique which worked without prior knowledge of training dataset, human intervention and linguistic knowledge [23]. This automated method increased the accuracy of baseline up to 76%.

Suresh and Raj [24] presented an aspect level method to find the sentiment of a particular brand using twitter feed with the help of novel fuzzy clustering and obtained accuracy of 76.4% with faster execution. Combinational effects of clustering were shown along with sentiment analysis on review datasets [25]. It was found that the K-means clustering algorithm provided better results than the balanced review datasets. The newly designed weighing system was shown to be better than traditional ones. Sentiments of movie reviews were analyzed by applying Word2Vec algorithm and K-means++ algorithm [26]. It was argued that this approach could be used for sarcasm and question detection with additional modifications.

A comparative study for sentiment analysis approach adopted by researchers has been shown in Table 1 by utilizing domain information and the languages for which the proposed work has been done is discussed. To extend the understanding, the stages of pre-processing has been taken into account followed by the machine learning techniques for analysis purpose and the accuracy that has been achieved so far.

Table 1. Comparative analysis of approaches adopted by researchers

Ref	Domain	Language	Preprocessing	Classification and clustering Algorithm	Limitations	Result
[27]	Labeled product review of four domains: books, DVDs, kitchen appliances, Electronics	English	1 Gram, 2 Gram, 1+2 Gram and 1 Gram +2 Gram approach	Classifier level fusion and feature approach	– The classifier level fusion faced unbalanced performance for multi-domain data.	80%
[8]	Blogs, forum text, review related to products	English, Dutch, French	Unigram, Bi Subjectivity, Bigram	Cascade learner classifier	– Lack of training data, – Conflicts sentiments, – lack of pattern detections	Approx 70%
[28]	Tweets which contain some noise, associated to the mobile operators	English	Normalized words	Annotated ensembles	– Word with different spelling and representation cannot be mapped	47% F-score
[26]	Movie review	English	Word2Vec	K-means/K-means++	– Couldn't enhance accuracy of baseline	-
[29]	Tweets	Spanish, English	Unigram, Bigram	Cascade Classifiers	– Normalization techniques and Slang dictionary can be include for Spanish	69% approx
[30]	Tweets on Indian political parties	Hindi	Hashtag, URL, Stopwords removal, negation handling	Dictionary based, naive Bayes and SVM algorithm	– Limited data size, – Emoticons are not include in data set – Only text data included	78%
[31]	Travel destination reviews	Hindi, Marathi	POS	SVM	missing concepts for Marathi language was there, by considering these the accuracy can be enhanced	72%
[32]	Social media text	Hinglish	Lowering Case, Lemmatization, Multiword Grouping	Convolution neural network (CNN), long short term frequency (LSTM), convolution neural network- bidirectional LSTM (CNN-BiLSTM)	Bilingual model required	83%

Researchers have worked on sentiment analysis in English as well as in native languages like Chinese, Spanish, Marathi, and Tamil. The work has been done in bilingual as well as multilingual like Hindi and Hindi-English combination, Tamilish, and English+French. There are many such combinations used in other pair of languages. However, combination of more than two languages has not been worked upon extensively. Due to the difficulties encountered for various reasons for example ambiguous words, inconsistent spelling, and part of speech and bag of words, sentiment analysis for the combination of Hindi-Hinglish-English languages have been missing from.

The researcher has mainly extended their hands on social network data, blogs, product reviews, and political reviews and subjective tweets. After the data collected by user from multiple sources, the pre-processing of data is done with the help of commonly used pre-processing techniques. As far as pre-processing steps are concern the most commonly used techniques are stop-word removal, lemmatization, stemming, lowering cases, part of speech, and normalization, which has shown the great effect on the accuracy of machine learning models.

The processed data is inputted to machine learning models for sentiment analysis purpose. For the classification and clustering methods, researchers have utilized from the simplest algorithms like naïve Bayes, SVM, K-means to the cascade classifiers and neural network techniques too. These techniques are mostly restricted themselves to the monolingual to bilingual concept, has taken into account the text data only and are restricted themselves with a small training dataset. Since the accuracy of these techniques is restricted to the range of 70-80%. So, the more chances are lies to enhance the accuracy level. Present paper is an attempt to highlight dataset creation from the tweets fetched from the twitter based on the hashtags. The pre-processing techniques utilized by researchers, the impact of these techniques and the comparative analysis of them are discussed in Table 1 (see in Appendix).

3. METHOD

To predict the human behaviour against an organization or entity or product, the sentiment analysis techniques play a vital role. In today's world where word-of-mouth, customer feedback, reviews and opinions have become major issues, sentiment analysis (SA) and opinion mining are the two techniques being used invariably [33], [34]. Subjective extraction of opinion related to an entity falls under opinion mining whereas is SA complete text analysis is performed [35]. SA represents sentiment identification in a text then followed by its analysis. The accuracy of decision making lies in the accuracy of sentiment analysis. The complete procedure of SA of multiple events, parallel running sub-event on social media (SM) and their influence on behaviour, reaction and even on thoughts of people have been discussed in [36]. The generalized process of performing SA from social networking data is as follows:

3.1. The opinion of users

The first step of SA is to retrieve the information (opinionative words, and phrases) from the huge amount of data available and to store this information in the required format [37]. Keeping main concern on finding influence of events and sub-events, data has been picked from twitter using the hashtag information to find events, user mentions and retweets to find the sub-event count. The information in the similar way is extracted and represented [38]. Figure 1 representing the word cloud for the extracted hashtags.

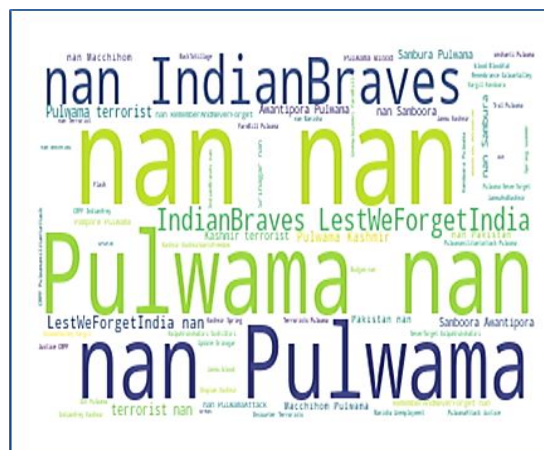


Figure 1. Wordcloud for extracted

3.2. Data cleaning

Data cleaning process is to make the text noise-free. The noise in the data is in the form of missing words or unwanted words. These missing or unwanted words are present in the disguise of some symbols which are not handled by the code. Data cleaning comprises information filtering, removal of stop-words/punctuation and tokenization process [36]. The Figure 2 given below provides the glimpse of the format of data, which is collected from twitter, is being cleaned after applying the clean text function. After the cleaning process data is converted into lowercase, tokenized and stopword have been removed and Figure 3 represents the process.

ip	created_at	source	original_text	clean_text
1.18E+18	Tue Oct 01	<a href="f	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE https://t.co/nBcYmVKmvE	THANK YOU FOR SUBSCRIBE
1.18E+18	Tue Oct 01	<a href="f	@isro @PMOIndia Hi Team, Since last two days, Have a chance to analyse the moon from surface of earth. It's reflect https://t.co/JDGB0JBjJS	Hi Team Since last two days Have chance analyse moon surface earth It's reflect
1.18E+18	Tue Oct 01	<a href="f	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE https://t.co/nBcYmVKmvE	THANK YOU FOR SUBSCRIBE
1.18E+18	Tue Oct 01	<a href="f	RT @kshamaLively: PM ko "Panauti" kehne aur mission ko fail kehne wale sune: USA (12 Attempts) Russia (7 Attempts) China (3 Attempts)	PM ko "Panauti" kehne aur mission ko fail kehne wale sune USA 12 Attempts Russia 7 Attempts China 3 Attempts To

Figure 2. Clean text hashtags

3.3. Data pre-processing

Data pre-processing consists of tokenization, part of speech, normalization, lemmatization and stemming of words where network among words is established [39]. Through this, all the relevant events and sub-events are mapped onto connecting words for example: -thanking you, thank you can be mapped onto thank. Here the stemming algorithm is applied on clean text. Figure 3 shows the stemmed dataset. The Algorithm 1 has been introduced to represent the overall procedure used for text processing.

Sentence ID	Token ID	Token
1	1	thank
1	2	you
1	4	subscribe
2	6	team
2	7	since
2	8	last
2	10	days
2	12	chance
2	13	analyse
2	14	moon
2	15	surface

Figure 3. Tokenized and stemmed data

Algorithm 1. Text preprocessing algorithm

```

for each tweet in Document do
perform tokenization by splitting the text
ignoring \. ' & ;'
end for
for each token in Document do
tweet.remove_stopwords(english)
tweet.remove_punctuations
tweet.remove_colon-symbol
tweet.replace_non_ASCII char with space
tweet.remove_emoticons.
end for
for each remaining word in dataset do
perform stemming using stemmer and store in Vector (Word List)
end for

```

3.4. Feature extraction

The feature extraction involves vectorization, bag of words, TF-IDF, N-Gram and word embedding techniques. The feature extraction maps data in vector space. In this phase, different hashtags which have similar meaning or which signify similar events are mapped together. For example, hashtags #chandrayan, #Chandrayan1, #ISRO, #IndiaFails signifies the same event 'chandrayan' and is clubbed together. Similarly, the tweets which have user-mentions similar to the events have also been clubbed together and considered as sub events.

3.5. Choosing machine learning techniques

The machine learning approaches based on problem definition are applied to find more accurate results. As per the discussion above, SVM and naïve Bayes classification techniques are mainly used when users have some predefined rules which are needed to be followed. For clustering techniques, K-means and Fuzzy logic algorithms are mainly employed due to their simplicity and accuracy [40]. However, some people have also used semi-supervised and hybrid techniques as well [20], [26].

3.6. Output for predictive level of sentiment analysis

Base on the predictive polarity levels, the polarity of the text is calculated [41]. Thereafter, on this calculated level of the polarity, sentiment is fixed as negative sentiment, positive sentiment or a neutral one. The process of performing sentiment analysis and then finding the polarity can be summarized as shown in Figure 4.

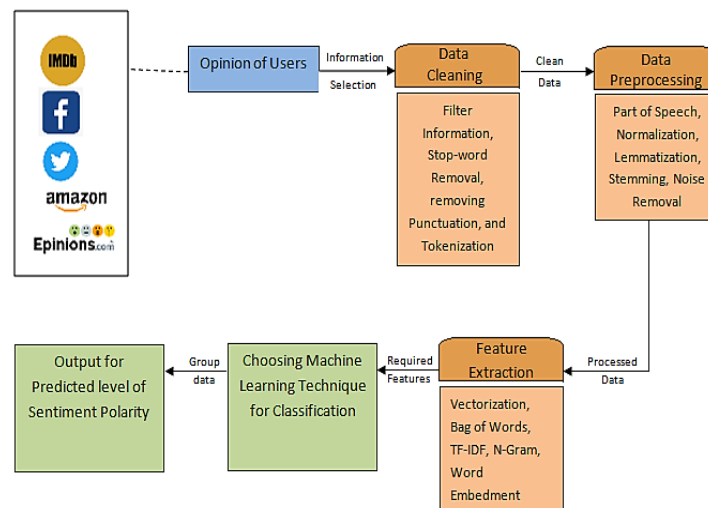


Figure 4. Method for sentiment analysis

4. RESULT

The results after applying each of pre-processing techniques namely tokenization, stopword removal and stemming have been shown in the Table 2 and plotted against the dataset size in Figure 5. Different values of dataset have been taken for testing the behaviour based on pre-processing stages. It has been

demonstrated that stopword removal and stemming are the compulsory parts for pre-processing. It has also been shown that the stopword removal has reduced the dimensionality of the text handsomely. From the Figure 5, it has been concluded that the application of pre-processing techniques has a positive impact on the number of terms selected. The results represent that negligible difference is shown in terms of numbers selected by stemming.

Table 2. Statistics for preprocessing

Dataset (Tweets)	Tokenization	Stop-word removal	Stemming
2,000	12,000	10,000	9,000
5,000	30,000	22,000	20,000
9,000	53,000	45,000	42,000
2,000	70,000	57,000	55,000

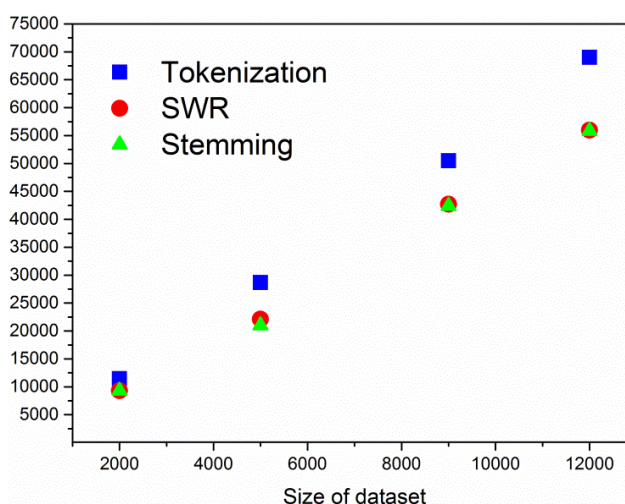


Figure 5. Effect of preprocessing

5. CONCLUSION




The present article discusses the supervised and unsupervised SA techniques. In this paper, the basic techniques of data extraction followed by the data cleaning and data pre-processing techniques have been presented. Three basic techniques for pre-processing i.e. tokenization, stopword removal and stemming have been introduced on twitter dataset. From the results, it can be concluded that pre-processing bears a huge impact to reduce the dimensionality of data which in-turn results in a high performing and more accurate SA techniques. The results prove that the stopword removal technique removes unnecessary words from the dataset and thereby improving accuracy. The same technique may be applied to the different dataset belonging to different domain. One can improve upon the list of stopword as per the domain and achieve better accuracy.

REFERENCES




- [1] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentimental reviews using machine learning techniques," *Procedia Computer Science*, vol. 57, pp. 821-829, 2015, doi: 10.1016/j.procs.2015.07.523.
- [2] T. Baldwin, P. Cook, M. Lui, A. Mackinlay, and L. Wang, "How noisy social media text, how diffrent social media sources?," *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 356-364.
- [3] A. Sarker and G. Gonzalez, "Data, tools and resources for mining social media drug chatter," *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, 2016, pp. 99-107.
- [4] C. S. P. Kumar and L. D. D. Babu, "Novel text preprocessing framework for sentiment analysis," *Smart Innovation, Systems and Technologies*, vol. 105, 2019, pp. 309-317, doi: 10.1007/978-981-13-1927-3_33.
- [5] G. Neha, "A review on the study of big data and big data analytics," *3rd International Conference on Computers and Management (ICCM 2017)*, 2017, pp. 276-285.
- [6] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial Intelligence Review*, vol. 52, pp. 1495-1545, 2019, doi: 10.1007/s10462-017-9599-6.
- [7] V. Sindhvani and P. Melville, "Document-word co-regularization for semi-supervised sentiment analysis," *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 1025-1030, doi: 10.1109/ICDM.2008.113.

- [8] B. Erik and M. M. Francine, "A machine learning approach to sentiment analysis in multilingual web text," *Information Retrieval*, vol. 12, no. 5, pp. 526-558, 2008, doi: 10.1007/s10791-008-9070-z.
- [9] J. Blitzer, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2006, pp. 440-447.
- [10] S. Li, C. R. Huang, G. Zhou, and S. Y. M. Lee, "Employing personal/impersonal views in supervised and semi-supervised sentiment classification," *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 414-423.
- [11] X. Zhu, "Semi-supervised learning literature survey contents," *Sci. York*, vol. 10, no. 1530, pp. 10, 2008, doi: 10.1.1.146.2352.
- [12] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification," *IJCAI International Joint Conference on Artificial Intelligence*, 2011, pp. 1826-1831, doi: 10.5591/978-1-57735-516-8/IJCAI11-306.
- [13] D. B. Liang PW, "Opinion mining on social media data," *2013 IEEE 14th International Conference on Mobile Data Management*, 2013, pp. 91-96, doi: 10.1109/MDM.2013.73.
- [14] P. H. Shahana and B. Omman, "Evaluation of features on sentimental analysis," *Procedia Computer Science*, vol. 46, pp. 1585-1592, 2015, doi: 10.1016/j.procs.2015.02.088.
- [15] T. K. Tran and T. T. Phan, "Deep learning application to ensemble learning-the simple, but effective, approach to sentiment classifying," *Applied Sciences*, vol. 9, no. 13, 2019, doi: 10.3390/app9132760.
- [16] F. Musumeci et al., "An overview on application of machine learning techniques in optical networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383-1408, 2018, doi: 10.1109/COMST.2018.2880039.
- [17] G. Li and F. Liu, "A clustering-based approach on sentiment analysis," *2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering*, 2010, pp. 331-337, doi: 10.1109/ISKE.2010.5680859.
- [18] M. Basavaraju and D. R. Prabhakar, "A novel method of spam mail detection using text based clustering approach," *International Journal of Computer Applications*, vol. 5, no. 4, pp. 15-25, 2010, doi: 10.5120/906-1283.
- [19] S. Venkatasubramanian, A. Veilumuthu, A. Krishnamurthy, C. E. V. Madhavan, K. Nath, and S. Arvindam, "A non-syntactic approach for text sentiment classification with stopwords," *Proceedings of the 20th International Conference Companion on World Wide Web*, 2011, pp. 137-138, doi: 10.1145/1963192.1963262.
- [20] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Clustering product features for opinion mining," *WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 347-354, doi: 10.1145/1935826.1935884.
- [21] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "A novel context-aware multimodal framework for persian sentiment analysis," *Neurocomputing*, vol. 457, pp. 377-388, 2021, doi: 10.1016/j.neucom.2021.02.020.
- [22] G. Li and F. Liu, "Application of a clustering method on sentiment analysis," *Journal of Information Science*, vol. 38, no. 2, pp. 127-139, 2012, doi: 10.1177/0165551511432670.
- [23] G. Li and F. Liu, "Sentiment analysis based on clustering: A framework in improving accuracy and recognizing neutral opinions," *Applied Intelligence volume*, vol. 40, no. 3, pp. 441-452, 2014, doi: 10.1007/s10489-013-0463-3.
- [24] H. Suresh and S. G. Raj, "An unsupervised fuzzy clustering method for twitter sentiment analysis," *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2016, pp. 80-85, doi: 10.1109/CSITSS.2016.7779444.
- [25] B. Ma, H. Yuan, and Y. Wu, "Exploring performance of clustering methods on document sentiment analysis," *Journal of Information Science*, vol. 43, no. 1, pp. 54-74, 2017, doi: 10.1177/0165551515617374.
- [26] K. Chakraborty, S. Bhattacharyya, R. Bag, and A. E. Hassanien, "Comparative sentiment analysis on a set of movie reviews using deep learning approach," *International Conference on Advanced Machine Learning Technologies and Applications*, vol. 723, 2018, pp. 311-318, doi: 10.1007/978-3-319-74690-6_31.
- [27] S. Li and C. Zong, "Multi-domain sentiment classification," *Proceedings of ACL-08: HLT, Short Papers*, 2008, pp. 257-260, doi: 10.3115/1557690.1557765.
- [28] A. Celikyilmaz, D. Hakkani-Tür, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," *2010 IEEE Spoken Language Technology Workshop*, 2010, pp. 79-84, doi: 10.1109/SLT.2010.5700826.
- [29] A. Balahur and J. M. P. Ortega, "Sentiment analysis system adaptation for multilingual processing: The case of tweets," *Information Processing & Management*, vol. 51, no. 4, pp. 547-556, 2015, doi: 10.1016/j.ipm.2014.10.004.
- [30] P. Sharma and T. S. Moh, "Prediction of Indian election using sentiment analysis on Hindi Twitter," *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1966-1971, doi: 10.1109/BigData.2016.7840818.
- [31] Balamurali AR, A. Joshi, and P. Bhattacharyya, "Cross-lingual sentiment analysis for indian languages using linked WordNets," *Proceedings of COLING 2012*, vol. 1, 2012, pp. 73-82.
- [32] T. T. Sasidhar, B. Premjith, and K. P. Soman, "Emotion detection in hinglish(hindi+english) code-mixed social media text," *Procedia Computer Science*, vol. 171, pp. 1346-1352, 2020, doi: 10.1016/j.procs.2020.04.144.
- [33] N. Rizk, A. Ebada, and E. Nasr, "Investigating mobile application Requirements evolution through sentiment analysis of users' reviews," *2015 11th International Computer Engineering Conference (ICENCO)*, 2015, pp. 123-130, doi: 10.1109/ICENCO.2015.7416336.
- [34] B. Liu, "Sentiment analysis and opinion mining," Morgan & Claypool Publishers, 2012.
- [35] M. Tsytsarou and T. Palpanas, "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery*, vol. 24, pp. 478-514, 2011, doi: 10.1007/s10618-011-0238-6.
- [36] N. Garg and K. Sharma, "Machine learning in text analysis, in handbook of research on emerging trends and applications of machine learning," *IGI Global*, 2020, pp. 383-402, doi: 10.4018/978-1-5225-9643-1.
- [37] C. L. Philip Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314-347, 2014, doi: 10.1016/j.ins.2014.01.015.
- [38] N. Garg and K. Sharma, "Sentiment analysis of events on social web," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 4, pp. 1232-1238, 2020, doi: 10.35940/ijitee.f3946.049620.
- [39] M. J. Denny and A. Spiriling, "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it," *Political Analysis*, vol. 26, no. 2, pp. 168-189, 2018, doi: 10.1017/pan.2017.44.
- [40] S. Yogesh, P. Bhatia, and O. Sangwan, "A review of studies on machine learning techniques," *International Journal of Computer Science and Security*, vol. 1, no. 1, pp. 70-84, 2007.
- [41] B. Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis, "Learning and predicting the evolution of social networks," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 26-35, 2010, doi: 10.1109/MIS.2010.91.

BIOGRAPHIES OF AUTHORS

Neha Garg    is currently working as an Assistant Professor, MRIIRS, Faridabad, India (around 10 years teaching experience), M.Tech from Banasthali Vidyapith, Near Niwai, Rajasthan and Ph.D. pursuing in computer science and engineering from MRIIRS, Faridabad, India. She has recently published a book “Analysis and Design of Algorithms- a Beginner Hope” with BPB Publication House. She has supervised and Guided research projects of B.Tech She have published research papers and book chapter in field of big data, and data mining. Her research interests are in the area of “Big Data Analytics” and “Machine Learning”. She can be contacted at email: gargsneha99@gmail.com.



Kamlesh Sharma    is currently working as an Associate Professor, MRIIRS, Faridabad, India (more than 14 years teaching experience). MCA, M.Tech from MDU University and Ph.D. in Computer Science and Engineering from Lingaya`s Vidyapeeth, India. Is currently Supervising five Ph.D. scholars. She has also supervised and guided research projects of M.Tech, B.Tech and application based projects for different competitions. She is also associated with four Govt. research projects in field of health recommender system, IOT, machine learning, AI and NLP. She has published more than 45 research papers in field of NLP, IOT, big data, green computing and data mining in reputed Journal (Web of Science, Scopus, UGC, Elsevier) and Conferences (ACM, IEEE). Her research area “Natural Language Processing” is based on innovative idea of reducing the mechanized efforts and adapting the software to Hindi dialect. She can be contacted at email: kamlesh.fet@mriu.edu.in.