

Sentiment Analysis of Tweets on Telangana State Government Flagship Schemes



K. Bhuvaneshwari, S A Jyothi Rani, V. V. Haragopal

Abstract: Over the last decade, the usage of social media has evolved to a greater extent. Today, social media platforms like Twitter, facebook, snapchat are vastly used to incept the opinions of public about a particular entity. Social media has become a great source of text data. Text analytics plays a crucial role on social media data to give answers to a wide variety of questions about public feedback on many issues or topics. The primary objective of this work is to analyse the public opinion or sentiment in social media on Telangana state government welfare schemes. The purpose of sentiment analysis is to find opinions from tweets and extract sentiments from them and find their polarity, i.e., positive, neutral or negative. Here we are using twitter as it has gained much popularity and media attention. The first step is to extract the tweets on particular schemes through Twitter API and Python language followed by cleaning and pre-processing steps of the raw tweets. Then tfidf vectoriser was invoked for feature extraction and creation of bag of words and finally sentiment polarity scores were obtained by using VADER (Valence Aware Dictionary and sEntiment Reasoner), lexicon and rule-based sentiment analysis tool.

Keywords: Sentiment Analysis, Twitter, Vader, Lexicon, Government Schemes

I. INTRODUCTION

In day to day life, social media is becoming an emerging sensation due to the expeditious development in information technology. People around the globe are using social media platform such as Twitter, Facebook, Instagram, snapchat and Tumblr etc on regular basis to incept their views or opinions about a wide variety of topics, social issues, products, and services, newly launched movies, government schemes, etc., which has made it a rich repository for text data. Twitter is one of the fast and simplest online platforms to express views towards an entity through short text called “tweets” [1]. When an individual shares his opinion or view on some emerging topic, feedback about a product, review of a movie or on any Social issue on social media website, other people around the globe also join in such conversations, and this forms a group of compatible users. These tweets can be analyzed to mine the real-time [2], [3] sentiments of the people about the topic discussed.

Manuscript received on 10 August 2022 | Revised Manuscript received on 29 August 2022 | Manuscript Accepted on 15 October 2022 | Manuscript published on 30 October 2022.

*Correspondence Author

K. Bhuvaneshwari*, Department of Statistics, Osmania University, Hyderabad, (Telangana), India. E-mail: k.bhuvaneshwari7@gmail.com

Dr. S. A Jyothi Rani, Department of Statistics, Osmania University, Hyderabad (Telangana), India.

Dr. V. V. Haragopal, Professor, Department of Statistics Mathematics, BITS Pilani, Hyderabad Campus, (Telangana), India.

©The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license: <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Sentiment Analysis (SA) is a Natural language processing technique deals with ex- traction of polarity i.e., positive, neutral or negative of an entity from a piece of text i.e., document or Phrase or a sentence. The concept of Sentiment analysis emerged in the 1990s, it has gained much attention since 2004 [4]. SA can be applied to any form of textual data of opinions such as reviews, blogs, and Microblogs like short text such as tweets. SA often helps any organization to know the opinions regarding any policy, product or service so that it guides for better decision making [5]. The same scenario can be applied in the government to aid the development and for making better decisions and fine-tune their policies by knowing the people’s attitude towards the policies or welfare programs being served. Opinion mining is a potent instrument that the government and its agencies can use to regularly gauge public opinion, which can pave the way for better governance in practical ways. In this work, we have proposed to perform opinion mining of public real-time tweets on various Telangana state government welfare programs using twitter dataset. Here we have collected tweets on various schemes of government starting from launch of the particular scheme till June 2022. The present study is the first approach under lexicon based sentiment analysis on Telangana State Government welfare schemes.

II. RELATED WORK

There has been a lot of recent study on sentiment analysis using lexicon-based [6] techniques, linguistic approaches, and other methods. The emotion of tweets is determined using linguistic features. The following is a list of earlier projects that have helped the area of sentiment analysis (SA) in recent years.

Wagh et al. [7] had developed a general sentiment classification model to use when the data is unlabeled the target domain. Gilbert [8] created the VADER technique, a straightforward rule-based model for general SA, and evaluated its performance against other common practise benchmarks, such as the General Inquirer, Senti Word Net, Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW), Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms. This study has demonstrated that VADER outperforms conventional approaches like LIWC in terms of advantages. Because VADER generalized more favorably to other domains and was more responsive to sentiment expressions in social media contexts than LIWC [9], it was separated from LIWC.



Sentiment Analysis of Tweets on Telangana State Government Flagship Schemes

The authors [10] focused analysing the pattern of the tweets on Swachh Bharat Abhiyan, a cleanliness campaign initiated by the Government of India by identifying the public opinion towards this initiative and determining the sentiment distribution of the mission region wise by considering the geolocations of the tweets. This analysis can be used to detect and apply approaches to make the mission gain popularity in the lesser known regions of the country.

The authors [11] opined that, the issue of automatically capturing emotion has two primary solutions, one method, Lexicon-based approach determines the orientation of a document based on the semantic orientation of its words or phrases. Building classifiers using labelled examples of texts or phrases is how the text classification approach works [12], which is effectively a supervised classification task. The second strategy is classified as a statistical or machine learning strategy.

In the proposed work, we extracted real-time tweets on welfare schemes of Telangana Government to find the polarity i.e., positive, negative or neutral through lexicon and rule based approach, i.e., VADER sentiment analysis particularly applicable to sentiments expressed in social media.

III. RESEARCH METHODOLOGY

The present study consists of 3 phases. Phase 1 is the acquisition of Twitter data on various government welfare programs. Phase 2 focuses on the pre-processing of the text i.e., raw tweets is carried out to purge and delete unnecessary content from the tweets. In phase three, the NLTK's VADER sentiment analyser is used to calculate the to obtain the sentiment scores and polarity of the schemes and finally visualization of the sentiment scores scheme wise, overall polarity values of all the schemes and sentiment distribution of various schemes year wise is done. The proposed step by step approach for this research study is given in fig 1

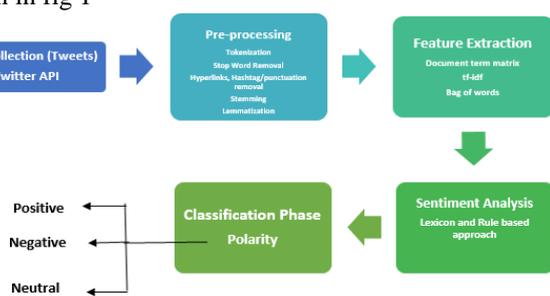


Fig. 1. Proposed Approach

A. Data Acquisition

Data collection is the primary step of Sentiment analysis [13]. In order to extract the required data i.e., tweets on particular government schemes, we have created a twitter developer account with the elevated access. Twitter allows to fetch the tweets by using Twitter API with the help of twitter credentials i.e., API key, API secret key, Access Token and Access Token Secret. In this work we used python programming language as it is open source and it consists of wide range of libraries for text mining. Using twitter API credentials i.e., keys and tokens in order to obtain twitter authorization to extract tweets, we collected

information such as authorid, id, date and time of tweet creation, text. Here we limited our search to English language. The data can be retrieved in JSON or CSV files. We collected tweets on various welfare schemes of Telangana state government right from launch of the scheme till June 2022, as in detailed in table 1.

Table- I: Scheme Wise Launch Dates and Tweets Extracted

S.No	Name of the Scheme	No. of Tweets Extracted	Scheme Launch date (dd-mm-yyyy)
1	Rythubandhu	50355	11-05-2018
2	Dalit bandhu	13978	03-08-2021
3	Mission bhagiratha	305022	06-08-2016
4	Kalyanlakshmi & shaadimubarak	3117	02-10-2014
5	Palle pragathi	7289	27-12-2014
6	Aasara pension	5007	08-11-2014
7	Harithaharam	64936	03-07-2015
8	RythuBhima	4978	14-08-2018

B. Data Pre-Processing

After data collection, the next primary step for any text mining application involves data pre-processing. The data in this study is tweets which is in text format that contains lot of noise like twitter handlers, hashtags, urls, emoticons and other irrelevant information. Hence before applying Natural language processing techniques, the repeated or duplicate tweets were removed and raw text was cleaned and pre-processed.

Pre-processing targets to process the tweets in an organized format which increases the machine understanding of the text [14]. The pre-processing steps includes. Lower case conversion in order to remove the difference between "Hello" and "hello", all the words are converted into lower case. Tokenization is segmenting the tweets into tokens (words) using delimiters, such as spaces and punctuations. All the tokenized words form a bag of words (BOW). Removing of stop words that carry no meaning and does not help in finding the sentiment of text such as "a", "an", "the", "have", "has", etc.

Links and hashtags must be removed as such links do not help in detecting Polarity. Punctuation marks also do not carry any meaning for the analysis and hence they can be detached from the text. Sometimes words which have similar roots but differ only in affixes, i.e., words are constant at the beginning and vary at the end. For eq., changed, changing, changes, changed have the same root change. Taking off the end part is called stemming and the residual part is called the stem. Lemmatization is the text normalization which involves changing the form of verb or adjective back to original form. For eq., the verb "studying is replaced with "study". Slang handling involves normalizing the unstructured tweets by eliminating the repeated characters. For eq., the word "gooooooooood" is cut short to "good".



C. Feature Extraction and Sentiment analysis

Feature Selection methods can be classified into lexicon-based methods and statistical methods [15] which are automatic methods that are most commonly used. Lexicon-based approaches usually begin with a small set of ‘seed’ words. The feature selection procedures treat the text either as Bag of Words (BOWs), or as a string which holds the sequence of terms in the text [16]. Bag of words is most frequently used approach because of its simplicity for the classification problems. The most commonly used feature selection step is the removal of stopwords and stemming.

Sentiment analysis (SA) generally referred as a method for extracting the subjectivity and polarity from the textual data and semantic orientation refers to assigning of sentiment values to the words and the terms which are semantically nearer to those words that gives the polarity and strength of words, phrases, or texts [17]. Here our interest is primarily to evaluate the semantics of the text.

Sentiment Classification practices can be broadly classified into machine learning approach, lexicon based approach and hybrid approach [18]. The Machine Learning Approach (ML) usually applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach depends on a sentiment lexicons, a collection of well-known and pre-compiled sentiment terms. It is divided into dictionary based approach and corpus based approach which uses statistical or semantic approaches to find the polarity. The hybrid Approach pools both the approaches.

In this analysis, we used lexicon and rule-based sentiment analysis tool i.e., VADER (Valence Aware Dictionary and s Entiment Reasoner). It is a combination of a sentiment lexicon, a list of lexical features which are labelled according to their semantic orientation as either positive or negative. It assigns the words to sentiment by considering the intensity of the word in the lexicon. This approach gives 4 sentiment components for each tweet. The first 3 components are positive, neutral, and negative. The last one is the normalization of all the previous mentioned 3 components of the tweet. The first three components will always sum upto one. These Compound score lies between -1 to +1 where -1 represents most negative and +1 denotes most positive sentiment.

For feature extraction, tfidf vectoriser is invoked and then sentiment polarity scores were obtained by applying NLTKs Vader module.

IV. RESULTS AND DISCUSSION

This section provides the detailed results of the study. Here we used twitter dataset on various Telangana state Government welfare schemes such as Rythubandhu, Dalitbandu, Mission Bhagiratha, palle Pragathi, Aasarapension, kalyanalakshmi & shaadi Mubarak, Rythu bhima and Haritha haram.

After preprocessing and cleaning of the raw text, we applied NLTKs Vader module from python programming language on each scheme. Here the metric metric Compound score is a value that calculates the sum of all the lexicon ratings that are normalized between -1(most extreme negative) and +1 (most extreme positive). Scheme wise polarity distribution of tweets is shown in fig 2 to fig.9.

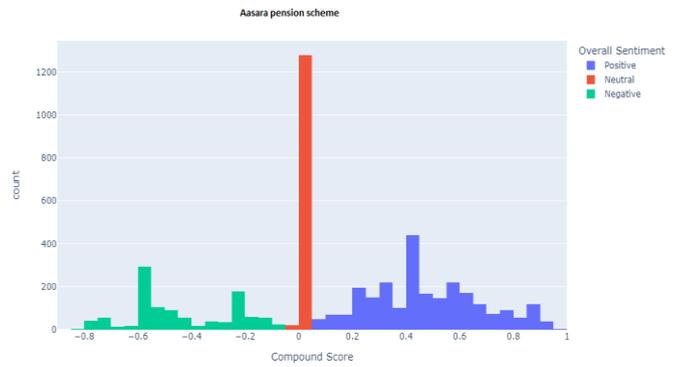


Fig. 2. Sentiment Distribution of Tweets -#Aasara Pension

Results of overall sentiments scheme wise are given below in Table 2.

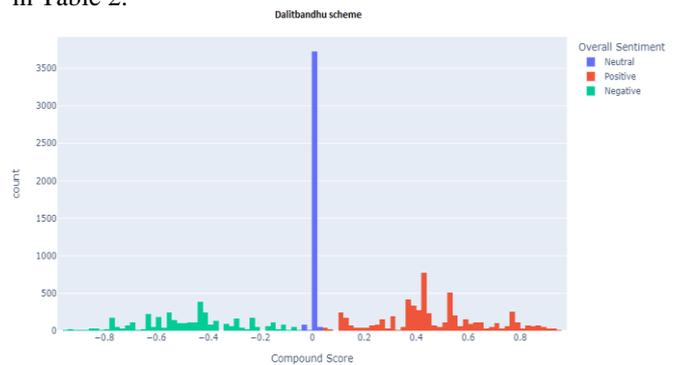


Fig. 3. Sentiment Distribution of Tweets - #Dalitbandu

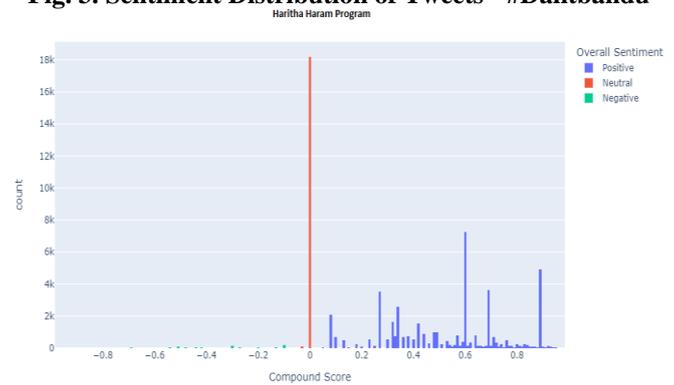


Fig. 4. Sentiment Distribution of Tweets - #Harithaharam

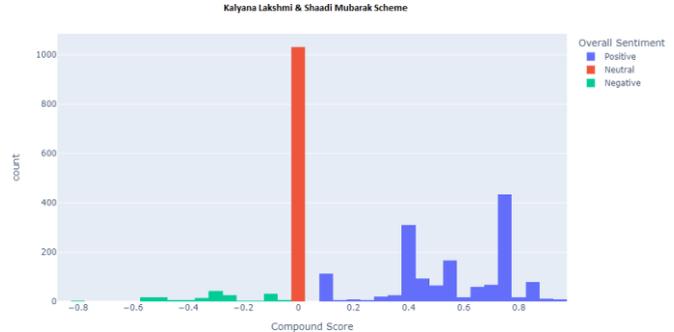


Fig. 5. Sentiment Distribution of Tweets - #Kalyanalakshmi & Shaadi Mubarak

Sentiment Analysis of Tweets on Telangana State Government Flagship Schemes

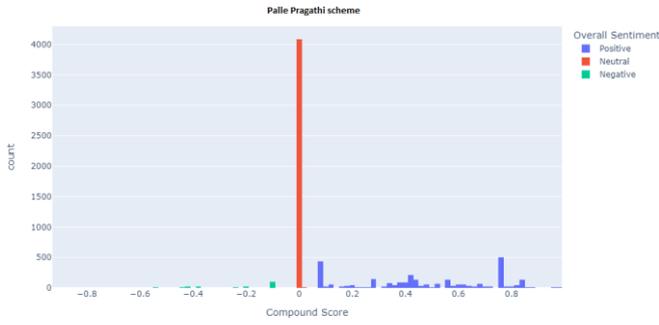


Fig. 6. Sentiment Distribution of Tweets - #Pallepragathi

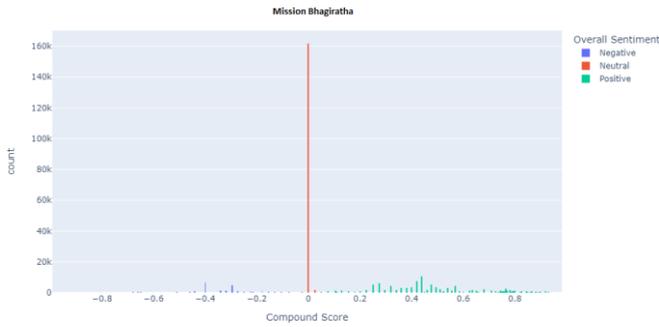


Fig.7.SentimentDistributionofTweets-#Missionbhagiratha

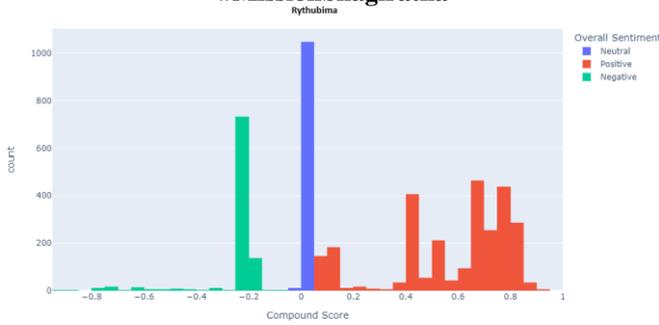


Fig. 8. Sentiment Distribution of Tweets -#Rythubima

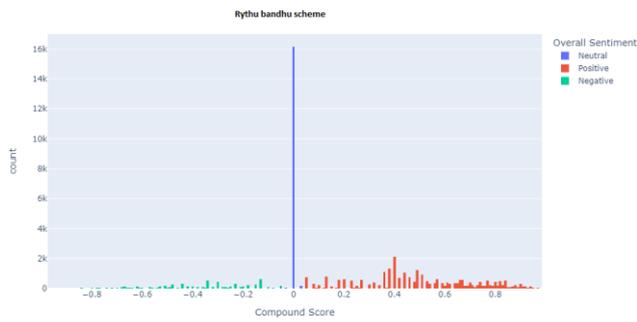


Fig. 9. Sentiment Distribution of Tweets -#Rythubandhu

Table- II: Scheme wise overall sentiment

Classification Scheme	Positive	Neutral	Negative	Total Tweets
#rythubandhu	26775	16446	6134	49355
#dalitbandhu	5769	3856	3715	13340
#kalyanakshmi & shaadimubarak	1507	1033	175	2715
#pallepragathi	4099	2899	232	7230
#harithaharam	44683	18306	1177	64166
#missionbagiratha	163986	113876	24975	302837
#aasarapension	2500	1297	1090	4887
#rythubima	2702	1059	973	4734



Fig. 10. Scheme wise Polarity Value

Fig.10, gives the detailed over view of the overall sentiments of all the Government schemes considered in this study. Among all the schemes, the positive sentiment is highest for Harithaharam program i.e, 69.64% followed by Rythubima (57.08%). Polarity results of tweets of the welfare schemes year wise starting from launch of the scheme are shown in fig 11.

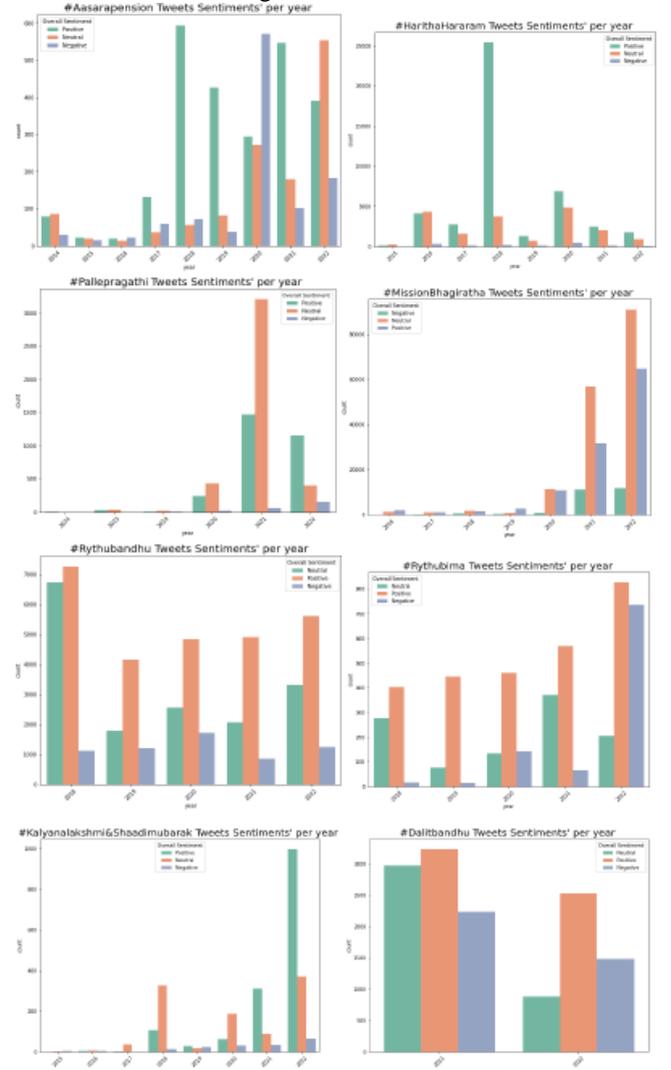


Fig. 11. Year wise Polarity of the Various Welfare Schemes



V. CONCLUSION AND FUTURE ENHANCEMENT

In this experimental research work, we used Twitter API and python which is open source. Tweets on various Telangana Government welfare schemes from twitter were collected and the text data was pre-processed using python libraries for sentiment analysis (SA) and opinion mining. Python is open source tool which is used in text mining applications and also to scrape the real-time streaming data from social media platforms like twitter and Facebook etc. There are many supervised and unsupervised methods through which Sentiment Analysis can be performed. In this study, as the dataset is unlabeled, we have used lexicon and rule based approach i.e., VADER (Valence Aware Dictionary and Sentiment Reasoner). This model has proven itself to be a powerful tool both in polarity and intensity of emotion when analyzing the sentiment on social media. In our analysis, sentiment scores for all the words in a tweet were calculated and then summed up which forms a metric called compound score which is normalized to be between -1 and +1. Based on this score, polarity of the tweets were identified. Further, the overall sentiment of various schemes were obtained given in fig.10. Also the fig.11 depicts the year wise sentiments of the public right from launch of the various welfare schemes. In future work, in addition to the implemented methodology, sentiment analysis of tweets in other languages apart from English can be of a great asset in opinion evaluation. Further, emoticons can also be included in this analysis for better ways of obtaining the polarity of tweets to have even more unerring results.

REFERENCES

1. A. Kumar and T. M. Sebastian, "Sentiment analysis on twitter," IJCSI International Journal of Computer Science Issues, vol. 9, no. 3, 2012, pp. 372–378.
2. V. Mittal and I. Kashyap, "Empirical Study of Impact of Various Concept Drifts in Data Stream Mining Methods, International Journal of Intelligent Systems and Applications, vol. 8, no. 12, 2016, pp. 65–65. [CrossRef]
3. V. Mittal and I. Kashyap, "Online methods of learning in occurrence of concept drift," International Journal of Computer Applications, vol. 117, no. 13, 2015. [CrossRef]
4. F. Gupta and S. Singal, "Sentiment analysis of the demonetization of the economy 2016, India region wise," IEEE 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence 2017, 2017, pp. 978–979. [CrossRef]
5. E. Cambria, D. Das, S. Bandyopadhyay, and Feraco, A practical guide to sentiment analysis, A., Ed., 2017. [CrossRef]
6. A. Vishal, Kharde, Prof, and Sheetal, "Sonawane "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, vol. 139, no. 11, 2016, pp. 5–15. [CrossRef]
7. C. J. H. E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," Eighth International Conference on Weblogs and Social Media.
8. B. Wagh, J. Shinde, and P. A. Kale, "A Twitter Sentiment Analysis Using Nltk and Machine Learning Techniques," Int. J. Emerg. Res. Manag. Technol, vol. 6, no. 12, 2018, pp. 37–44. [CrossRef]
9. S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, 2009.
10. S. Raj and T. Kajla, "Sentiment analysis of Swachh Bharat Abhiyan," Int. J. Bus. Analytics Intell, vol. 3, no. 1, 2015, pp. 2321–1857. [CrossRef]
11. P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in Proceedings of 40th Meeting of the Association for Computational Linguistics, 2002, pp. 417–424. [CrossRef]
12. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79–86. [CrossRef]

13. T. Wilson, J. Wiebe, and P. Hoffman, "Recognizing contextual polarity in phrase-level sentiment analysis," Proceedings of HLT/EMNLP, 2005. [CrossRef]
14. W. Y. Chong, B. Selvaretnam, and L. K. Soon, "Natural language processing for sentiment analysis," IEEE 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology.
15. Y. Mejova and P. Srinivasan, "Exploring feature definition and selection for sentiment classifiers," Proceedings of the fifth international AAAI conference on weblogs and social media, 2011.
16. A. Charu, C, and Z. C. Xiang, "Mining Text Data," LLC, vol. 12, 2012.
17. Y. Mejova and P. Srinivasan, "Exploring feature definition and selection for sentiment classifiers," Proceedings of the fifth international AAAI conference on weblogs and social media, 2011.
18. D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," Proceedings of the 8th international conference on the semantic web, ESWC'11, 2011, pp. 88–99. [CrossRef]

AUTHORS PROFILE



K. Bhuvaneshwari, is a Research scholar in Department of Statistics, Osmania University, Hyderabad, Telangana, India. Her Research interest includes Text analytics and Natural language processing.



Dr. S. A. Jyothi Rani, is Associate Professor in Department of Statistics, Osmania University, Hyderabad, India. Her area of research includes Time series analysis, Machine learning, Deep learning and AI.



Dr. V. V. Haragopal, is Retired professor from Department of Statistics, Osmania University, Hyderabad, India. Presently he is working as Professor of Statistics, Department of mathematics, BITSPilani, Hyderabad Campus, India. His area of research includes Operations Research, Statistical Pattern Recognition, Group Testing, Regression Analysis and Multivariate DataAnalysis, Text mining and NLP. Also involved in many data analysis projects.