International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Dynamic Data Analysis and Decision Making on Twitter Data

Nikitha Kumari, Prabhakar kandukuri

Abstract: It became a tedious task for the data analysts to make decisions on social networks. The existing approaches are not adequate to perform data pre-processing, analysis and decision making on the data dynamically. Therefore, this research aims to propose an approach to data analysis and decision making. The proposed approach emphasis on extracting tweets form twitter API (Application Program Interface), pre-processing the tweets by following seven pre-processing steps. The processed tweets are trained by NLTK (Natural Language Toolkit) and Text Blob are given to the sentiment analysis. Classification is done using the Naive Bayes algorithm to make a decision on processed tweets. The tweets which are related to "MeToo Movement" are considered primarily for decision making and satisfactory results are obtained. It is been observed that the proposed approach is accurate when compared to other approaches.

Keywords: Twitter, Text pre-processing, Machine learning, Sentiment Analysis, MeToo movement.

I. INTRODUCTION

Nowadays social media has become a most important media for the source of communication and gaining information about the latest news and events which are being posted on social media where billions of active users share their opinion according to their interest such as sports, education, and transport research area, medical and health, politics, share market and many other activities. So, by collecting their posts or a message, tweets can be performing a sentiment analysis to the tweets. Sentiment analysis where we can find whether the tweets are positive tweets, negative tweets or the neutral tweets.

Sentiment analysis is otherwise also called as "opinion mining" or "feeling artificial intelligence" and insinuates the usage of common language handling Natural Language Processing (NLP), content mining, computational etymology, and bio estimations to systematically perceive, remove,

assess, and look at enthusiastic states and emotional data [20]. Similarly, the proposed approach performing sentiment analysis in this paper to the tweets related to Metoo movement.

Revised Manuscript Received on November 30, 2019.
* Correspondence Author

Nikitha Kumari*, Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India. Email: nikithakumari408@gmail.com.

Prabhakar kandukuri, Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India. Email: prabhakarcs@vardhaman.org

1.1 Metoo Movement

The Metoo development, with a huge assortment of the neighbourhood and global elective names, is a development against inappropriate behaviour and rape. Metoo development in India is an appearance of the universal. MeToo development that happened in late 2018 (as a result to show day) in parts of Indian culture including government, media, and the Bollywood film industry. In India, the MeToo development is viewed as either an autonomous outgrowth affected by the universal crusade against lewd behaviour of ladies in the working environment. Me Too started picking up unmistakable quality in India with the expanding prominence of the worldwide development, and later assembled sharp energy in October 2018 in media outlets of Bollywood [16]. And the billions of tweets were posted about these #metoo movement in the twitter account. This research is considered with a widely used social networking site which is Twitter. By collecting the tweets related to the metoo movement sentiment analysis is been performed on the tweets to find the polarity of the tweets. This paper is been organize as follows: section 2 consists of the background work. Section 3 describes details about the steps involved in the proposed work. Section 4 consists of and result & discussion. Section 5 concludes the work by throwing a lightened on to the future direction.

II. RELATED WORK

This part of section is related with several related works on the sentiment analysis on the twitter data are discussed. Symeon, Dimitrios, and AviArampatzis had compared sixteen pre-processing techniques on two datasets for analysis and by using four algorithms of machine learning such as Bernoulli Naive Bayes (BNB), Linear SVC, Convolutional neural network (NCC) and supply regression [1] they distinguished performance classes based on the result and the ablation study it shows on analysis some techniques give good result in classification whereas the techniques like marking up capitalized words, replacing slang, removing punctuation, replacing negation with antonyms, and such as spelling correction decrease an accuracy. Zhao and Gui had examined results of pre-processing techniques on twitter knowledge for the defense wall of sentiment classification [2] they had collected the tweets that carries with its symbols, abbreviation, folksonomy and uncorrected words. By removing the URLs, user mentions, hashtags, punctuation, and importance of slang words, spelling correction was identified. SVM was used to carry out the experiment.

Tejinder and Madhu has focused on the effects of pre-processing techniques by using six pre-processing techniques with five twitter datasets and two models of features and 4 four classifiers by increasing the strategies of replacing negations and expanding acronym, small by removing URLs, stop words and numbers [3]. Ahmad, Elang, Wisnu, Akbar, and Ridwan had collected the data from many Indonesian official twitter accounts which is always been updated regarding football news. LDA is been used for topic modelling and obtained many topics such as pre-match analysis, football club achievements, and live match's update [4].

According Prerna, Ranjana and Pankaj opinion mining has turned into a developing theme of research because of a great deal of stubborn information accessible on social networking site. In [5] they have collected the tweets and performed a sentiment analysis on the prime minister Modii's digital India companion by using the dictionary-based approach they had obtained the percentage of (50,20,30) positive, negative and the neutral tweets respectively. Trupthi, Suresh and Narasimha have performed an ongoing nostalgic investigation on the tweets that are extricated from the twitter and gives the time sensitive investigation to the client in [6] experiment is done by using the Hadoop and map reduce which is a trained module and classification is been done by the Naive Bayes.

Yuling and Zhi zhang had consolidated the upsides of CNN and SVM and builds a content notion examination model dependent on CNN and SVM by using more trained word vector as an output in [7] the content powerful investigation model dependent on CNNs and SVM proposed in their paper can viably improve the exhibition of the content characterization. Kavya and Narasinga had built up a model which performs assumption examination on Twitter information utilizing Machine Learning Method is Naive Bayes algorithm and the model which was proposed in their exploration was utilizing the Hadoop structure for handling the dataset [8].

Huma and Shikha had conducted an experiment by collecting the tweets from the twitter by using the Hadoop framework for processing the movie dataset.by using these tweets they had clean the tweets and classified the tweets using the machine learning algorithm that is Naive Bayes algorithm and compared the tweets with the emoticons a without the emoticons, they had got the better results by considering the emoticons [9].

Ali, Sana, Ahmad and Shahaboddin had conducted an experiment on the twitter data by gathering the tweets related to the political reviews based on these reviews the sentiment analysis is done to the tweets by using the Naive Bayes algorithm and SVM and had also provided a comparison of these two techniques [10].

Imane, Rochdi, Alexis and Abdessamad has gathered a tweet related to the US election which was held in 2016 in a large number of volume and stored in the HDFS and by building dictionaries and the classification of the data is been done, data is been pre-processed and the sentiment analysis is carried out to the data by using the Naïve Bayes algorithm [11]. Ankita and Anand conducted a sentiment analysis on the review provided by the customers during their flight gathered an US airline tweets from dataset called Kaggle dataset for six major airlines and the sentiment analysis are done by using seven different machine learning classification [12].

Lowri, Christian, Michael, Alun and Irena have shown the estimation of figures of speech as highlights of feeling examination by demonstrating that colloquialism-based highlights essentially improve opinion characterization results when such highlights are available. The general execution as far as accuracy, review, and F-measure was improved from 45% to 64% in one examination, and from 46% to 61% in the other. By gathering the informational index of 580 figures of speech that are applicable to sentiment analysis [17].

Hemant, Jyotim, Bhagyashri and Ganesh had done the sentiment analysis proposed a methodology, that it's simple and better to do the conclusions investigation of twitter information utilizing Hadoop and enormous information with Naïve Bayes Algorithm [18]. Sahar, Feddah and Wejdan they had generated a module that shows the sentiment analysis results on the basis of the data collected on two topics such as KFC and McDonalds to show the most popular restaurant by using the different machine learning algorithms [19].

In the previous research work we noticed that the sentiment analysis and the pre-processing is performed on the data related to the tweets such as football matches, airlines, and the twitter data but this research mainly focus on the tweets related to the trendy topic in the twitter which is "Metoo Movement".

III. DYNAMIC DATA ANALYSIS

This section is designed as the part of the proposed work where these following steps are been carried out such as fetching the tweets, processing the tweets, and analyzing the data for the classification as shown in fig. 1.

3.1 Fetching the tweets

To collect the tweets from the twitter API from the well-known famous social media site i.e. Twitter, Firstly we need to connect to the Twitter API account, while creating an account in Twitter API we get a consumer key, consumer secret, access token and the access token secret key, by accessing this key an connection will be established and the tweets are been collected which is related to the Metoo Movement.

3.2 Processing the Tweets

After fetching the tweets from the Twitter API. Generally, tweets consist of many spelling mistakes and many short forms in the sentences and exclamation marks with multiple question marks in the sentences so the process of pre-processing techniques is been used to clean the tweets. The process of cleaning or filtering the tweets is known as the processing of the tweets. which is done in the next process by applying the pre-processing techniques to get the processed tweets. Pre-processing techniques are as follows:

i)Removes Hashtags in front of words: A hashtag is a keyword or an expression used to depict a subject or a topic, which is quickly gone before by the pound sign (#). ... Twitter clients put hashtags in their tweets [21]**.** By removing the hashtags tweets are been cleaned by removing an extra hashtag.

ii)Removing Unicode string: Sometimes tweets are also consisting of a non-English term which are been removed by this process, the Unicode strings such as "\u002c" and "x96" contains in the tweets.

iii)Removing Stop Words: Most commonly used words in the languages such as (‘a’, ‘but’, ‘an’, ‘in’, ‘and’, ‘at’) are stop words. It can also be said as the sentences which contain a word with high frequency, which is not useless in the sentences is been removed.

iv)Removing Multiple Exclamation Marks, Question marks and Punctuation Marks: Some sentences in tweets consists of many multiple exclamation’s marks and the extra unwanted question marks which will be removed by this process.

Punctuation such as (‘!’, ‘.’, ‘comma (,)’, ‘?’). The sentences which consists of punctuation marks that denotes the existences of few sentiments which suggests associate intensive positive or the negative sentiment, therefore when we have a tender to take away punctuation marks it’ll decrease the efficiency [1].

v)Spelling Corrections: Some of the tweets consisting of spelling mistakes in the tweets by the process the spellings mistakes are been corrected.

vi)Removes Emoticons: Emoticon is a portrayal of a facial expression, for example, a grin(smile) or glare, framed by different mixes of console characters and used to pass on the tweeters emotions or expected tone by this process an emotions symbols will be removed from the text.

vii)Replacing Slang and Abbreviations: Tweets written by the users contains many short forms and slang. Slang may be a kind of language consists of phrases and also the words. shortening of the words are known as abbreviations. Some of the examples ‘omg’, ‘TBT’, ‘ty’, which are phrases and are commonly used while tweeting which literally means ‘oh my god’, ‘Throwback Thursday’ and ‘thank you’.

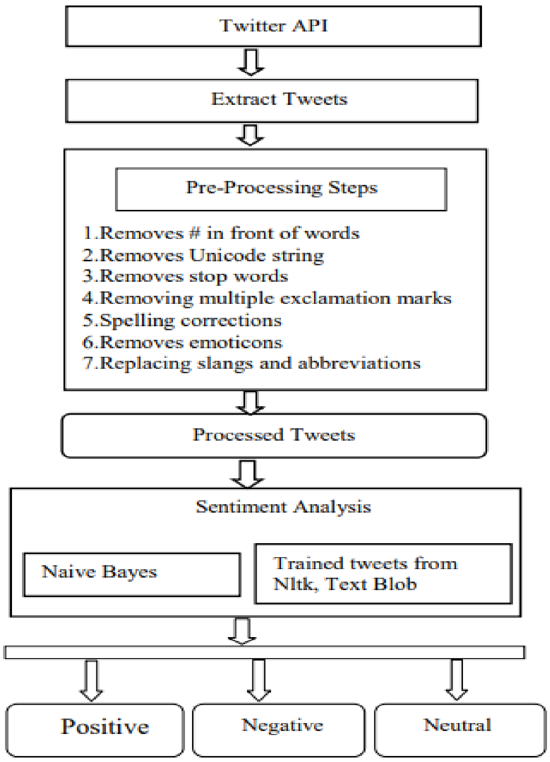

Fig.1: Dynamic data analysis and decision-making approach

3.3 Sentiment Analysis

By cleaning the tweets with all the pre-processing techniques, the processed tweet is been obtained and classification is done by the methods such as Naive Bayes algorithm, NLTK, and the text blob classifiers are applied on the processed tweets to obtained a tweets analysis such as a positive tweet, negative tweets, and neutral tweets. The following methods of classifiers are discussed below

i) Naive Bayes Algorithm

Naïve Bayes calculation is a valuable procedure to apply in text classification. The primary point of utilizing this calculation is a result of its incredibly quick identified with other classification techniques. This Classification is named as Naive Bayes after Thomas Bayes, who proposed the Bayes Theorem of deciding likelihood which decides accurate probabilities for speculation and furthermore it is hearty to clamour in info information [9].

$$P(C/X) = \frac{P(X/C), P(C)}{P(X)}$$

Where P (C | X) is back likelihood, P (X | C) is probability, P(C) is class earlier likelihood and P(X) is the predictor earlier probability.

ii) NLTK

Nltk a library in python, which is one of the most dominant NLP libraries which contains bundles to cause machines to comprehend human language and answer to it with a suitable reaction. Tokenization, Stemming, Lemmatization, Punctuation, Character check, word tally is a portion of these bundles [13]. Nltk which also gives the base to substance getting ready and gathering Errands, for instance, tokenization, naming, isolating, content control can be performed with the usage of Nltk. These libraries furthermore embody diverse trainable classifiers by using Naïve Bayes Classifier [15].

In this paper, Nltk is been used in a classifier by providing the trained tweets to obtain the polarity of the tweets. After the step of obtaining the processed tweets is been done by the Nltk and the Text blob along with the Naïve Bayes classifier.

iii) Text blob

Text blob is a Python library for preparing literary information. It gives a straightforward API to plunging into regular NLP assignments, for example, grammatical feature labelling, thing phrase extraction, slant examination, characterization, interpretation [14]. In this paper by using the trained tweets of Text blob analysis is done on the processed tweets to obtain a polarity of the tweets.

IV. RESULTS AND DISCUSSIONS

Results are been obtained by following the procedure which is been discussed in the proposed method by extracting the real-time tweets from the twitter API and apply the pre-processing techniques such as removing the stop words, removing # tags from the tweets, correcting them, removing punctuation marks, replacing the short forms and removing the emoticons from the tweets by applying these techniques the tweets are been cleaned.

After these processes the processed tweets are analyzed by using the NLTK and the Text blob and the Naive Bayes algorithm is used after the analyzing the results obtained is the percentage of the positive tweets, negative tweets and the percentage of the neutral tweets and also printing the first five positive and negative tweets as shown in the sample output fig:1.

A. Processed Tweets and Tweets before the pre-processing

Table1 shows the difference in the tweets after applying the pre-processing techniques to the tweets. Commas, full stops, spelling corrections, punctuation marks and the short forms are been removed from the tweets.

Table 1 Processed Tweets and Normal Tweets

Normal Tweets	After pre-processing
indeed 98 Ishiro Moral of the stories #Metoomovement is cancerous, @reed_indeed_98 @J_Ishiro Moral of the story Metoo movement is cancerous	indeed 98 Ishiro Moral of the story Metoo movement is cancerous reed indeed 98 J Ishiro Moral of the story Metoo movement is cancerous
And they are doing round two with the Metoo movement People don't understand that ALL human beings.	And they are doing round two with the Metoo movement People dont understand that ALL human beings
Many male comedians argue with the way women approach the metoo movement what's one.	Many male comedians argue with the way women approach the metoo movement whats one
Kelly Dodd Someone needs to re-educate those women on the MeToo movement They won't.	Kelly Dodd Someone needs to reeducate those women on the MeToo movement They wont
Definitely a pervert who thinks the metoo movement has gone too far and he can't compliment be a skee.	Definitely a pervert who thinks the metoo movement has gone too far and he cant compliment be a skee

B. Table Showing the Polarity of the Tweets

Table2 describes about the polarity of the tweets whether it is positive, negative or the neutral on the basis of their polarity rating it is been classified. These polarities of the tweets are obtained while fetching the tweets, an excel file is created and loaded all the tweets into the file.

Table 2: Showing the Polarity of the Tweets

Tweets	Polarity
Quite an interesting study So the MeToo movement seems to have instead of causing more awareness amp emancipation f	Positive
Does he understand that being a victim of the MeToo movement means he assaulted or harassed a woman	Negative
Kelly Dodd Someone needs to re-educate those women on the MeToo movement They won t	Neutral
The new Dave Chappelle stand up was hilarious He made fun of white's blacks Asians the LGBT community poor people	Positive
New study reveals the MeToo movement has backfired Those are steps backward	Positive
metoo movement is completely different	Neutral
Oh, I was naming amp shaming long before the MeToo movement took its first steps	Positive
You know I realized those headings of like omg men are scared to shake a female's hand because of the metoo movement	Neutral
Whatever 0 0001 chance you had for a metoo movement for the gaming industry is now dead So is your career	Negative
The MeToo movement isn t cancel culture It is what BECAME cancel culture as the hashtag it	Neutral

C. Results and Analysis

Fig.2 is an output which is obtained by following the proposed approach as the steps discussed in the fig1. Where it got the percentage of 57.50 positive tweets, 22.50 of negative tweets and the 20 percent of neutral tweets in the result.

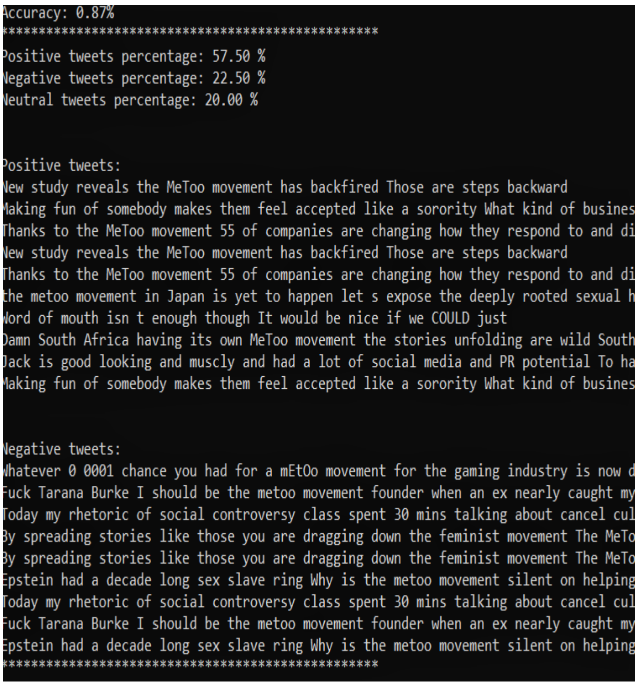

Fig.2: Decision on dynamic data

Retrieval Number: A5255119119/2019©BEIESP
DOI: 10.35940/ijitee.A5255.119119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

D. Graph for Percentage of the Tweets

Fig.3 shows the percentage of the tweet obtained after the analysis done the results obtained such as the 57.50% of the tweets are positive about the MeToo movement, 22.50% tweets obtained is negative and the 20% of tweets are negative.

According to the Bhumika and Monika [15] they had got a 66.24 percent of accuracy by using the Naïve Bayes algorithm. Where they had done a survey on their paper on the sentiment analysis by using the machine learning algorithm on python. And also, Abdullah and Zubair khan in [20] had surveyed the sentiment analysis on twitter data where they had concluded that by using the Naïve Bayes algorithm had an accuracy of app 80 percent of accuracy.

The proposed approach describes an experiment with sentiment analysis of the twitter data which is related to the "Metoo Movement", the proposed approach got the polarity of the tweets such as positive tweets, negative tweets, and the neutral tweets and got the accuracy of 87% obtained by measuring the polarity of the tweets and how accurate the result is obtained in an approach.

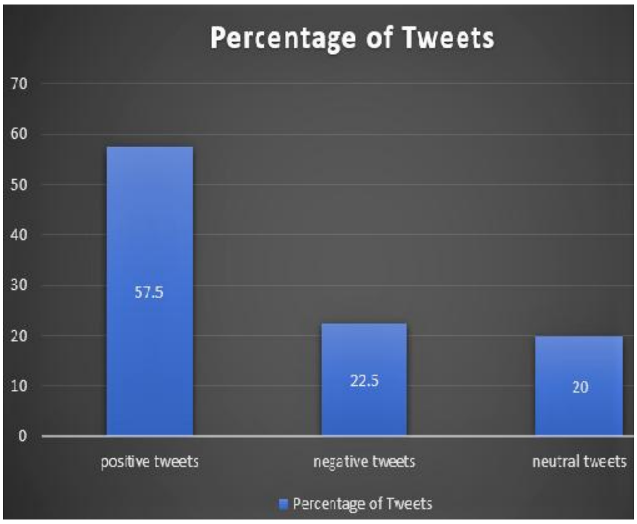

Fig.3: Dynamic data analysis

V. CONCLUSION

The proposed work describes a holistic approach which presents an experiment on sentiment analysis of the twitter data i.e. related to the Metoo movement is been extracted from the twitter a social networking site and the tweets are been processed and send to the classification. Tweets are built and trained by using the NLTK and text blob and the Classification is done by using the Naive Bayes algorithm where results are obtaining such as the positive, negative and neutral tweets respectively. By using the proposed approach has archived an accuracy of 86% by measuring the polarity rate of the tweets. which is been considered as good accuracy. When the tweets are written in the native language then the need of an approach is been needed to convert the native to the English tweets this can be considered as a future work. And model can additionally upgrade to any ideal level if one needs to by fusing more highlights in the database.

REFERENCE

1. SymeonSymeonidis, Dimitrios Effrosynidis, and AviArampatzis: "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis" in Expert Systems with Applications Elsevier, pp:298-310, 2018.
2. Zhao Jianqlang and GuiXiaolin: "comparison research on text pre-processing methods on twitter sentiment analysis" IEEE access, 5, 2870-2879, 2017.
3. Tajinder and Madhu Kumari: "Role of text pre-processing in twitter sentiment analysis" procedia computer science 89 (2016) 549-554.
4. Ahmad FathanHidayatullah, ElangCergasPembrani, Wisnu Kurniawan, Gilang Akbar, Ridwan Pranata: "Twitter topic modelling on football news". International conference on computer and communication systems,2018 in Indonesia.
5. Prerna Mishra, Ranjana Rajnish and Pankaj Kumar: "Sentiment analysis of twitter data: case study on digital India" International conference on information technology in 2016.
6. M.Trupthi, Suresh Pabboju and G. Narasimha: "Sentiment analysis on twitter using streaming API", International advance computing conference in 2017.
7. Yuling chen and Zhi zhang: "Research on text sentiment analysis based on CNN and SVM", 13th IEEE conference on industrial electronics and applications on 31may-2 June 2018.
8. Kavya Suppala and Narasinga rao: "Sentiment analysis using Naïve Bayes classifiers", international journal of innovative technology and exploring engineering June 2019.
9. Huma Parveen and Shikha Pandey: "Sentiment analysis on twitter data-set using Naive Bayes algorithm", 2016 IEEE 2nd international conference on applied and theoretical computer and communication technology.
10. Ali Hasan, Sana Moin, Ahmad Karim and Shahaboddin Shamshirband "Machine learning-based sentiment analysis for twitter accounts", MDPI, 2018.
11. Imane el Alaouli,Youssef Gahi, Rochdi Messoussi, youness chaabi, Alexis tadoskoff and Abdessamad Kobi "A novel adaptable approach for sentient analysis on big social data" Springer 2018.
12. Ankita Rane and Anand Kumar: "Sentiment classification system of twitter data for US airlines services analysis", 2018 42nd IEEE international conference of computer and software and applications.
13. Guru tutorial point "Guru99 Tech Pvt Ltd",Copyright - Guru99 2019,https://www.guru99.com/nltk-tutorial.html
14. Text blob https://textblob.readthedocs.io/en/dev/
15. Bhumika Gupta and Monika Negi: "Study of twitter sentiment analysis using machine learning algorithms on python" international journal of computer applications (0975-8887) volume 165-no.9, may 2017.
16. Wikipidiahttps://en.wikipedia.org/wiki/Me_Too_movement_(India).
17. Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece and Irena Spasic: "The Role of Idioms in Sentiment Analysis", *Expert Systems with Applications*, Elsevier, (2015).
18. Hemant J. Kamble, Jyotim Ingale, Bhagyashri R. Posture and Ganesh S. Ghuge "Sentiment Analysis of Twitter User Using Bigdata and Hadoop", published in International Journal of Engineering Research in Computer Science and Engineering, v-5 I-3 2018.
19. Sahar A El Rahman, Feddah Alhumaidi and Wejdan Abdullah Alsehri: "Sentiment analysis on twitter data" IEEE 2019.
20. Abdullah alsaeedi and mohammad Zubair khan: "A study on sentiment analysis techniques of twitter data", international journal of advanced computer science and applications, vol-10, no.2,2019.
21. Google search https://www.google.com/search?

AUTHORS PROFILE

Nikitha Kumari pursuing her MTech in computer science and engineering at Vardhaman college of engineering, Hyderabad, India. She received her B. tech degree in computer science and engineering from Aurobindo college of engineering Ibrahimpatnam, India. Her main reaserch interest includes data science and cloud computing. She has published a paper in UGC approval journal.

Dr. Prabhakar Kundukuri received his Ph. D. in Computer Science & Engineering at JNT University Anantapur, Anatapuramu, Andhra Pradesh, India. He received his M. Tech. Degree in Computer Science & Engineering from JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India, He received his B. Tech. Degree in Computer Science & Engineering from Acharya Nagarjuna University, Guntur, Andhra Pradesh, and He received his Diploma in Computer Engineering (D.CM.E) from the State Board of Technical Education and Training, Hyderabad, erstwhile Andhra Pradesh.

He is an Associate professor of Computer Science & Engineering Department, Vardhaman college of Engineering, Hyderabad, India. His main research interest includes Software Engineering, data Science and Web Services. He published several papers in various international journals/ conferences. Member IAENG.