iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

Fake News Detection and Sentiment Analysis in Twitter

Gayatri Potey[1], Rucha Jadhav[2], Kushagra Shroff[3], Anish Gore[4], Mrs. Dhanashree Phalke[5], Mr. Jayant Shimpi[6]
[1, 2, 3, 4, 5, 6]*Department of Computer Engineering, D Y Patil College of Engineering Akurdi, Pune, Savitribai Phule Pune University, Pune, India*

Abstract: ***Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Online networking is producing a huge measure of slant rich information as tweets, notices, blog entries and so on. Sentiment analysis allows businesses to identify customer sentiment toward products, brands or services in online conversations and feedback. Twitter sentiment analysis is difficult and different when compared to general sentiment analysis due to the presence of slang words and misspellings. Information base methodology and Machine learning approach are the two procedures utilized for examining notions from the content. Open and private conclusion about a wide assortment of subjects are communicated and spread persistently by means of various web based life. Twitter is one of the online networking that is picking up notoriety. Twitter offers associations a quick and compelling approach to examine clients' points of view toward the basic to achievement in the commercial centre. Developing a program for sentiment analysis is an approach to be used for computationally measuring customers' perceptions. This project uses knowledge base including various patterns for tweets along with multiple strategies to detect the sentiment expressed in a tweet and if a tweet is genuine or not. Various machine learning and knowledge base approaches are used to compare patterns and apply strategies and NLP for sentiment analysis.***

Keywords: NLP (Natural Language Processing), Sentiment Analysis, Machine Learning, Pattern Matching, Twitter Data, POS (Part of Speech)

I. INTRODUCTION

Twitter has risen as a significant small scale blogging site, having more than 100 million clients creating more than 500 million tweets each day. Because of essence of such enormous crowd, Twitter reliably draws in clients to pass on their sentiments and point of view about any issue, brand, organization or some other subject of intrigue. Because of this explanation, Twitter is utilized as an instructive source by numerous associations, establishments and organizations. On Twitter, clients are permitted to impart their insights as tweets, utilizing just 140 characters. This prompts individuals to compact their announcements by utilizing slang, contractions, emoticons, short structures and so on. Alongside this, individuals pass on their feelings by utilizing mockery and polysemy. Thus it is supported to term that the Twitter language is unstructured.

So as to extricate notion from tweets, conclusion examination is utilized. The outcomes from this can be utilized in numerous territories like dissecting and checking changes of notion with an occasion, notions in regards to a specific brand or arrival of a specific item, examining general visibility of government strategies and so forth. A great deal of research has been done on Twitter information so as to arrange the tweets and break down the outcomes. In this task the point is to anticipate the assessments from tweets by checking the extremity of tweets as positive, negative or insignificant. Assumption examination is a procedure of determining conclusion of a specific proclamation or sentence.

It's a characterization strategy which gets assessment from the tweets and figures a feeling and based on which, estimation order is performed. Slants are abstract to the subject of intrigue. It is required to detail the sort of highlights that will be chosen for the assessment it encapsulates. In the programming model, feeling alluded to, is the class of substances that the individual performing slant examination needs to discover in the tweets.

The element of the feeling class is critical factor in choosing the proficiency of the model. For instance, there can be two-class tweet opinion arrangement (positive and negative) or three class tweet conclusion order (positive, negative and immaterial). Notion investigation approaches can be comprehensively sorted in two classes – dictionary based and AI based. Vocabulary based methodology is solo as it proposes to perform examination utilizing dictionaries and a scoring technique to assess sentiments. While AI approach includes utilization of highlight extraction and preparing the model utilizing highlight set and some dataset. The essential strides for performing opinion examination incorporates information assortment, pre-handling of information, include

extraction, choosing standard highlights, assumption discovery and performing arrangement either utilizing straightforward calculation or, in all likelihood AI draws near. Highlights resemble Parts-of discourse highlights for example things, qualifiers, descriptors, and so on in every tweet are labeled. For the exactness motivation behind extremity location if slants alongside different information base examples and numerous AI methodologies are utilized to assess the conclusions. Nearby the validity of the tweet will be checked or not or if has impacted by different tweets which can be valuable in bits of gossip moderation on social medias. This methodology will create the higher precision for extremity by considering POS factor and validity also as can be utilized in different areas, for example, breaking down item surveys or government strategies, and so on where it very well may be found if adverse impact is spread and in the event that it influences individuals.

II. LITERATURE SURVEY

In the paper [1], Beakcheol Jang conducted comprehensive measurements to understand the characteristics, including similarities and differences, of data from the news and SNSs. The observed differences are as follows: It is challenging to find the same topic in the news and SNS.

The news responds to official events whereas SNSs respond to personal interests. The news mentions a specific topic continually, whereas the transition from one topic to another in SNSs is fast. The issues discussed on SNSs are different every day. The news can identify specific events with a single keyword, but many keywords are required to find the required data in SNSs.

In the paper [2] proposed a new method for the calculation of polarities and strengths of Chinese sentiment phrases is which could be used to analyse semantic fuzziness of Chinese. It uses a probability value, rather than_xed value forth polarity strengths of sentiment phrases, compared with the conventional methods.

In project [3], a new approach for sentiment analysis is proposed by MONDHER BOUAZIZI AND TOMOAKI OHTSUKI in 2017, where a set of tweets is to be classified into 7 different classes. The obtained results show some potential: the accuracy obtained for multi-class sentiment analysis in the data set used was 60.2%. However, a more optimized training set would present better performances.

[4]The project proposed a sentiment analysis method for news based on a linear regression model. The method developed by Aldo Hernández Victor Sanchez in 2016 employs natural language processing analysis on a collected corpus and determines negative sentiments within a specific context.

The objective is to predict the response of specific groups involved in hacking activism when the sentiment is negative enough among different News users.

In [5] (Exploring Sentiment Analysis on News Data) the author Manju Venugopalan and Deepa Gupta proposed a hybrid news sentiment classification model incorporating domain oriented lexicons, unigrams and news specific features using machine learning techniques has been developed and the classification accuracies have been found to improve by an average of around 2 points across different domains.

The effectiveness of incorporating concepts of domain specificity in the polarity lexicon and the capacities of explicit news features to extract sentiment has been validated. Pruning the unigrams based on their significant presence in a class has simplified the model to a large extent.

In [6] project, - Rincy Jose and Varghese S Chooralil have implemented a real time, domain independent twitter sentiment analyser using sentiment lexicons such as SentiWordNet and WordNet. It compared political sentiment towards two politicians by plotting graphs using results of sentiment analysis on real-time extracted twitter data. This was done by applying WSD and negation handling in order to increase accuracy of sentiment analysis. Negation handling results in 1% improvement in classification accuracy and WSD results in 2.6% improvement in classification accuracy.

III.EXISTING SYSTEM

In Existing System, to analyse the behaviour of news required maximum resources. To analyse the fake news, we required man power to deep down into it and check the authentication of news. All possible connection with news has to be checked manually. It is time consuming and costly approach. Limitation of existing system:-

A. Time Consuming Process
B. Man-Power Required.
C. Deep Knowledge required.
D. Cost driven approach.

IV. PROPOSED SYSTEM

In the proposed system, tweets will be retrieved from twitter using twitter API based on the query. The collected tweets will be subjected to pre-processing. The various patterns and strategic algorithms including some of machine learning algorithms for NLP to supervise the data will be then applied. The results of the algorithms i.e. the sentiment and influence will be represented in graphical manner (pie charts/bar charts). The proposed system is more effective than the existing one.

This is done to know how the statistics determined from the representation of the result can have an impact in a particular field as well as influence of negativity spread by rumours.

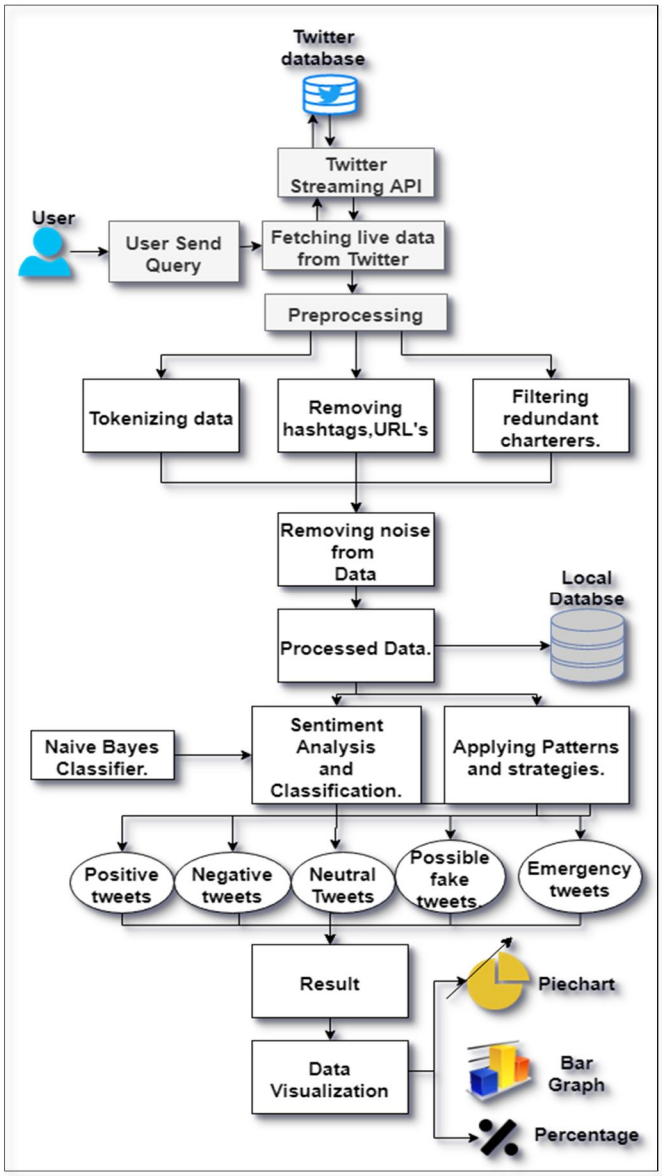


Fig.1:- System Architecture

User fetches the tweets from twitter which is the crude information for the framework. The tweets are fetched from the twitter databases which are accessed by the program with the help of a twitter developer account key. These tweets experience information pre-handling where in the hash labels and URL's are expelled from the tweets since tweets are commonly arranged after the hash label framework and furthermore experiences with expulsion of repetitive characters. Here tokenization of information additionally happens, where a token is a bit of an entire, so a word is a token in a sentence, and a sentence is a token in a section. Here tokenization is the way toward parting a string into a rundown of tokens. This whole pre-handling assists with cleaning, parse and tokenize the tweets.

Later, the noise which is the stop words are identified and removed from the data. This processed data is sent to the local database where patterns and strategies are applied. Naïve Bayes Classifier is applied for sentiment analysis to the processed data. The yield of this is the information gets separated into parameters of positive tweets, negative tweets, unbiased tweets, conceivable phony tweets and crisis tweets. This gives the conclusive outcome for Fake news recognition and assessment examination, which is additionally shown as pie charts, bar graphs and percentage.

V. METHODOLOGY (NATURAL LANGUAGE PROCESSING)

A. *Fake News/Rumour Detection*

A rumour can be described as a temporal communication network, where each node communicates to the user, the edges correspond to communication between nodes and the temporal aspect captures the propagation of messages through the network. The intuition is that there are measurable differences between the temporal communication network corresponding to false and true rumours. In order to capture these differences, characteristics of rumours need to be identified. It makes sense that these characteristics would be related to either the nodes (i.e. users) in the network, the edges (i.e. messages) in the network or the temporal behaviour of the network (i.e. propagation).

B. *Linguistic*

The linguistic features capture the characteristics of the text of the tweets in a rumour. A total of 4 linguistic features were found to significantly contribute to the outcome of our models. In the descending order of contribution these features are: ratio of tweet containing negations, average formality & sophistication of the tweets, ratio of tweets containing opinion & insight, and ratio of inferring & tentative tweets. The features are described in detail.

1) *Vulgarity:* The presence of vulgar words in the tweet.
2) *Abbreviations:* The presence of abbreviations (such as b4 for before, ‘JK’ for just kidding and ‘IRL’ for in real life) in the tweet.
3) *Emoticons:* The presence of emoticons in the tweet.
4) *Average Word Complexity:* Average length of words in the tweet
5) *Sentence Complexity:* The grammatical complexity of the tweet.

C. *User Identities*

The user features capture the characteristics of the users involved in spreading a rumour. A total of 6 user features were found to significantly contribute to the outcome of our models. In the descending order of contribution these features are: controversialist, originality, credibility, influence, role, and engagement. We will now describe each of these features in detail.

D. *Propagation Dynamics*

The propagation features capture the temporal diffusion dynamics of a rumour. A total of 7 propagation features were found to significantly contribute to the outcome of our models. In the descending order of contribution these features are: fraction of low-to-high diffusion, fraction of nodes in largest connected component (LCC), average depth to breadth ratio, ratio of new users, ratio of original tweets, fraction of tweets containing outside links, and the fraction of isolated nodes. All of these features are derived from a rumour’s diffusion graph. Before describing these features in detail, diffusion graph creation process needs to be understood.

E. *Sentiment Analysis*

Sentiment Analysis is the procedure of 'computationally' deciding if a bit of composing is sure, negative or impartial. It's otherwise called sentiment mining, inferring the conclusion or demeanour of a speaker.

F. *TextBlob*

TextBlob is a Python (2 and 3) library for preparing literary information. It gives a straightforward API to jumping into normal regular language handling (NLP) undertakings, for example, grammatical form labelling, thing phrase extraction, estimation investigation, order, interpretation, and that's only the tip of the iceberg. TextBlob is a python library and offers a basic API to get to its strategies and perform essential NLP assignments.

Something to be thankful for about TextBlob is that they are much the same as python strings. Along these lines, you can change and play with it same like done in python.

VI. DATASET

The dataset used here is live data/tweets from twitter which are fetched at runtime and provide with real time tweets. These are tweets which are authentic tweets from real users on twitter.

The dataset used is dataset from the NLTK (Natural Language Toolkit) package used in python where 7000 positive tweets and 7000 negative tweets of the dataset are split into training and testing in the ratio of 10:3. The corpus from the NLTK package is also used as a dataset for stop words, positive words and negative words.

VII. MATHEMATICAL MODEL

Let 'S' be the system

Where,

S= {I, O, P, Fs, Ss}

Where,

I = Set of input

O = Set of output

P = Set of technical processes

Fs = Set of Failure state

Ss = Set of Success state

Identify the input data I1, I2,...,In

I = {(Twitter Data)}

Identify the output applications as O1, O2,...,On

O = {(Rumors Detection, Fake News Detection, Sentiment Detection)}

Identify the Process as P

P = {(Data Processing, Natural Language Processing, Sentiment Analysis, Pattern Recognition)}

Identify the Failure state as Fs

Fs = {(If fake news not predicted}

Identify the Success state as Ss

P = {(Fake news detected successfully)}

VIII. RESULT & DISCUSSION

In proposed system we have created one web based application using Python's Flask framework which is light weight. In proposed application we are fetching real time tweets from twitter data and applying algorithm on them to get result out of that. To access data from twitter, you need to have authenticated twitter developer account which allows you to access the data. After accessing the data we are also storing that data into SQL database. Then we applied algorithms on that data. For sentiment analysis, we are using Textblob and NLTK libraries. And for fake news detection, we have used TFIDF algorithm. It's taking approximately two to five minutes for execution. According to network complexity it fetches tweets from server and processes them accordingly. Following screenshot shows the final output of project which has interpreted COVID19 keyword on twitter and which is accurate and efficient.

Fig.2:- Fake News Detection

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

IX. CONCLUSION

This project set out to solve a practical problem of sentiment analysis and genuinely check of Twitter posts. We proposed a method using knowledge base patterns, strategies and machine learning approaches. These methods are proposed to increase the accuracy of sentiment check for tweets. Patterns can be used to evaluate if the tweets was a influenced rumour or a genuine post by any user. By using API of twitter it is possible to work on live tweets than to work on offline data. Querying and fetching of particular tweets from twitter is possible by using its API. Finding influence or negativity spread by users can be useful in various analytical tasks.

X. FUTURE SCOPE

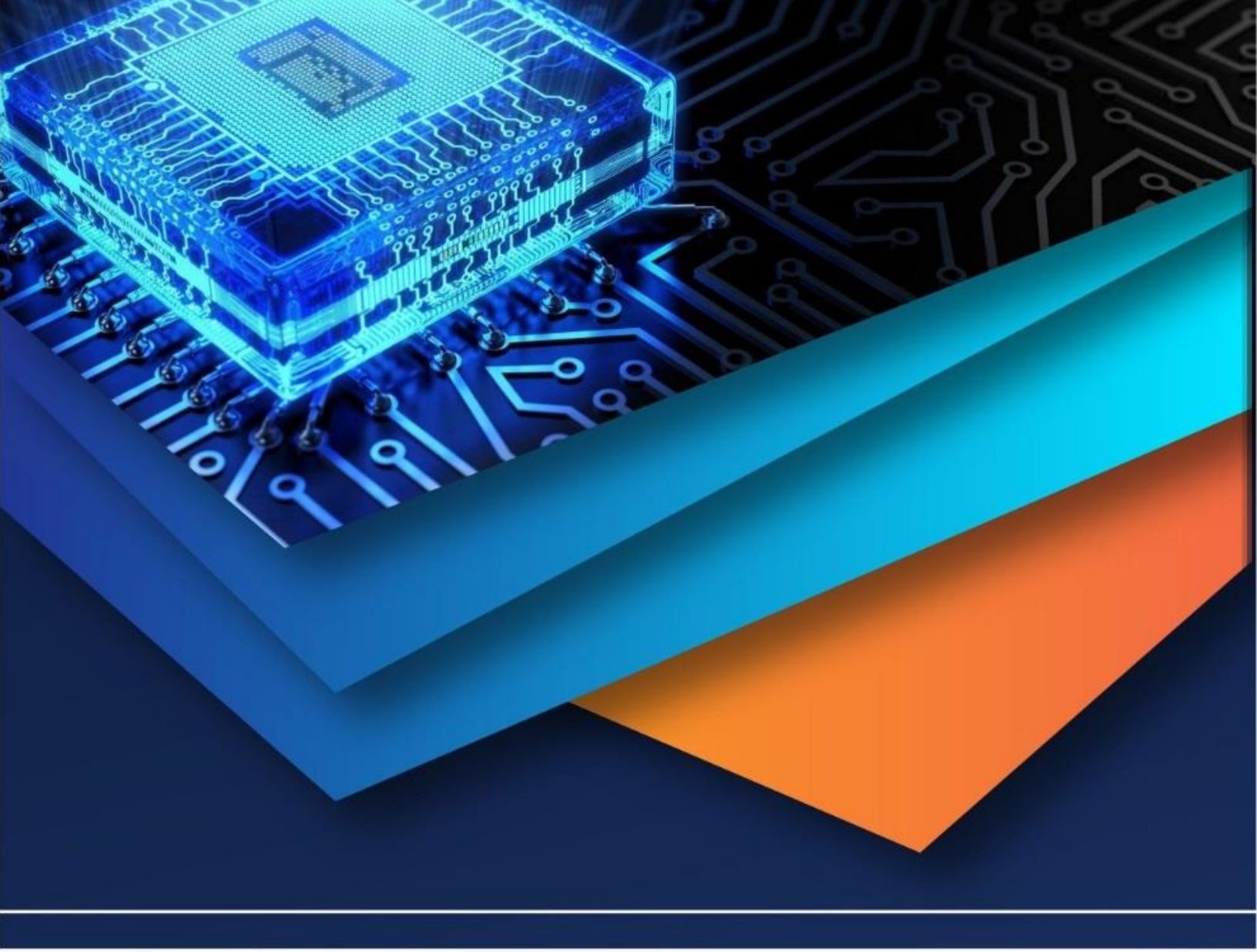
In future we can integrate this system with other social media platform. We can optimize time complexity by using better resources. We can add more parameter to enhance the accuracy of system. We can implement this system with several other algorithms to compare the accuracy.

XI. ACKNOWLEDGEMENT

We wish to express our profound thanks to all who helped us directly or indirectly in making this paper. We am especially grateful to our guide Mrs. Dhanashree Phalke and co-guide Mr. Jayant Shimpi for their time to time, very much needed, and valuable guidance. Without their full support and cheerful encouragement, the paper would not have been completed on time.

REFERENCE

[1] Beakcheol Jang and Jungwon Yoon "Characteristics Analysis of Data from News and Social Network Services" DOI 10.1109/ACCESS.2018.2818792, IEEE.
[2] HAI TAN AND JUN ZHANG "Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness" 2169-3536 2018 IEEE.
[3] MONDHER BOUAZIZI AND TOMOAKI OHTSUKI, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter" 2169-3536 2017 IEEE.
[4] Aldo Hernández, Victor Sanchez "Security Attack Prediction Based on User Sentiment Analysis of Twitter Data" 978-1-4673-8075-1/16/$31.00 2016 IEEE
[5] Manju Venugopalan and Deepa Gupta "Exploring Sentiment Analysis on Twitter Data" 978-1-4673-7948-9/15/$31.00 ©2015 IEEE
[6] Rincy Jose and Varghese S Chooralil "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation" 978-1-4673-7349-4/15/$31.00 ©2015 IEEE
[7] Anurag P. Jain and Mr. Vijay D. Katkar "Sentiments Analysis Of Twitter Data Using Data Mining" 978-1-4673-7758-4/15/$31.00 ©2015 IEEE
[8] Gaurav D Rajurkar and Rajeshwari M Goudar "A speedy data uploading approach for Twitter Trend And Sentiment Analysis using HADOOP" 978-1-4799-6892-3/15 $31.00 © 2015 IEEE



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY