iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4335

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: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Importance of Machine Learning in the Growth of Twitter

Kritika Pandey[1], Rupal Jain[2], Mrs. Shruti Ahuja[3]
[1, 2]*Students of Mahavir Swami Institute of Technology, B.Tech, CSE, Guru Gobind Singh Indraprastha University, Delhi, India*
[3]*Head of Department of Computer Science Engineering, Mahavir Swami Institute of Technology, Guru Gobind Singh Indraprastha University, Delhi, India*

Abstract: Twitter is a widely used platform by millions of user to express their opinions, suggestions and feelings on different occasions, which is becoming popular from the past decades. To process and analyse the data from Twitter, called tweets, analysis methods such as sentiment analysis and topic modelling are used. There is a different format of the tweets, that is, it is of limited words which produces new problems like use of slang, abbreviations etc. This paper contributes to this analysis. Firstly, we will apply pre-processing steps to the tweets to extract features from it. Then, a machine learning based approach is proposed to estimate the sentiments of a tweet. This method is required to extract topics from the training dataset, and train models for each of those topics. This method allows to increase the accuracy of the sentiment estimation as compared to a single model for every topic.
Keywords-component: Twitter; sentiment analysis; opinion mining; natural language processing; feature extraction; topic modelling

I. INTRODUCTION

Twitter is emerging as a major micro-blogging social website which allows its users to share short messages called Tweets which can have only 140 characters at maximum. It has over 100 million users which generates over 500 million tweets every day. Twitter is being used as an informative source by many organizations, institutions and companies. The post words limit leads to compacting of statements by using slang, abbreviations, emoticons, short forms etc. Along with this, people convey their opinions by using sarcasm. Now we can say that it is justified to term the Twitter language as an unstructured language.

The arrival of these networking sites like twitter, which gives real-time information, have developed the creation of an unequalled collection of opinions about every global entity that is of interest. Twitter poses newer and different challenges because it is an excellent channel for opinion creation and presentation. Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. That is why it is possible to collect text all users post from different social and interests groups.

According to this study, social media are growing more and more popular and they are now one of the fastest way of communication for both people and companies. Twitter's slogan is: "Twitter it's what's happening". A lot of research has been done on Twitter data so as to classify the tweets and analyse the results. Several research projects have applied sentiment analysis to twitter date in order to extract general public opinion regarding any political issues. This paper proposes a sentiment analysis method: First, topics are extracted from the training dataset. Then an algorithm is trained for each topic. Finally, the method classifies the sentiment of a sentence according the best topic related algorithm results.

A. Problem Statement

There are many software to extract data regarding a person's sentiment on a specific product or service, organizations and other data, but the workers are still facing issues regarding the data extraction. Some issues are such as:

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

Difficulty of Sentiment Analysis with inappropriate English

Informal language refers to the use of everyday language and slang in communication, employing the conventions of spoken language such as 'would not' and 'wouldn't'. But not all systems are able to detect sentiment from use of informal language and this could hanker the analysis and decision making process. Short-form is widely used even with short message service (SMS). This is done so as to minimize the characters used.

Performing sentiment analysis is a challenging task on Twitter data, as we mentioned earlier. Here we define the reasons for this:

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Limited size of tweets Usage of slang Using Twitter features like use of hashtags, user reference and URLs. Users express their opinions in a variety of ways, using different language in between, or using repeated words or symbols to convey an emotion.All these problems are required to be faced in the pre-processing section. Apart from these, we are facing problems in feature extraction having less features in hand and reducing the dimensionality of features.

B. Objective

The aim while performing sentiment analysis on tweets is basically to classify the tweets in different sentiment classes accurately. In this field of research, various approaches have evolved, which proposes methods to train a model and then test it to check its efficiency. In this paper we aim to review of some researches in this domain and study how to perform sentiment analysis on Twitter data using Python.

II. RELATED WORK

The first approach for extracting sentiments from texts was human generated baseline but this method was not able to handle the complexity of the language, and was providing low accuracy results. Another most popular approach for sentiment analysis is supervised machine learning based techniques. The three main machine learning algorithms which are applied for sentiment analysis are: Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM). The accuracy of these three algorithm depends on the feature extraction method which is applied and on the analysed datasets. For example, SVM gives better performance when it uses only unigrams, but adding bigrams features will reduce its accuracy. The feature extraction methods have to take some specificity into account: Presence is better than frequency; Negative handling; Bigrams; part of speech tags; Lemmatizing or streaming word.

A. Sentimental Analysis

Sentiment analysis is a process of deriving sentiment of a particular statement or sentence. It is a classification technique which intends to comprehend the opinions derived from the tweets, formulate them and distribute them into the categories like positive, negative, neutral. In the programming model, sentiment is the class of entities through which we can classify the tweets. For deciding the efficiency of the model, a crucial factor known as dimension of the sentiment class is used. For example, in two-class tweet sentiment classification we can have positive and negative entities, or in three class tweet sentiment classification, we can have positive, negative and neutral entities.

Sentiment analysis approaches can be broadly categorized in two classes:

1) *Lexicon Based Approach:* It is unsupervised approach as it proposes to perform analysis using lexicons and using a scoring method to evaluate opinions.
2) *Machine Learning Based:* It involves use of feature extraction and training the model using feature set and some dataset.

The basic steps for performing sentiment analysis includes data collection, pre-processing of data, feature extraction, selecting baseline features, sentiment detection and performing classification either using simple computation or machine learning approaches.

The Sentiment Analysis tasks can be done at several levels of granularity: Word level
Phrase or sentence level Document level
Feature level.

The methods of automatically annotating sentiment at the word level fall into the following two categories: (1) dictionary-based approaches and (2) corpus-based approaches. Further, to automate sentiment analysis, different approaches have been applied to predict the sentiments of words, expressions or documents. These include Natural Language Processing (NLP) and Machine Learning (ML) algorithms.

Sentiment analysis on Twitter posts is the next step in the field of sentiment analysis, as tweets give us a richer and more varied resource of opinions and sentiments about anything. Tweets have many unique characteristics:

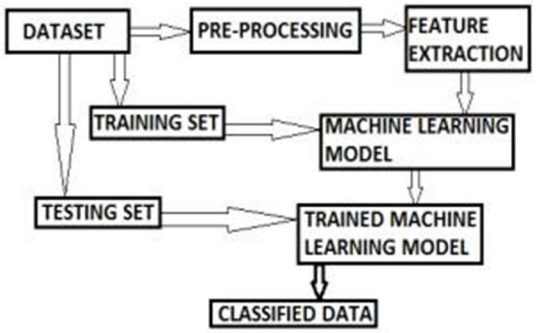
3) *Message Length:* The maximum length of a Twitter message is 140 characters.
4) *Writing Technique:* The occurrence of incorrect spellings and cyber slang in tweets, apart from which messages are quick and short, people use acronyms, misspell, and use emoticons and other characters that convey special meanings.
5) *Availability:* The amount of data available is immense. The Twitter API facilitates collection of tweets for training.
6) *Topics:* Twitter users post messages about a range of topics.
7) *Real Time:* Tweets are being limited to 140 letters and are updated very often. This gives a more real time feel and represents the first reactions to events while blogs are longer and take time.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

III. METHODOLOGY

This project has been divided into 2 phases. First, literature study is conducted, followed by system development. Literature study involves conducting studies on various sentiment analysis techniques and method that currently in used. In phase 2, application requirements and functionalities are defined prior to its development. Also, architecture and interface design of the program and how it will interact are also identified. In developing the Twitter Sentiment Analysis application, several tools are utilized, such as Python Shell 2.7.2 and Notepad.



IV. LITERATURE REVIEW

A. Tweet Collection

It refers to the broad area of natural language processing, text mining, computational linguistics, which involves the computational study of sentiments, opinions and emotions expressed in text.

Tweet analysis is mostly used in application domains including accounting, law, research, entertainment, education, technology, politics, and marketing. In earlier days many social media have given web Users Avenue for opening up to express and share their thoughts and opinions. Tweet collection involves gathering relevant tweets about the particular area of interest. The tweets are collected using Twitter's streaming API or any other mining tool for example WEKA. The dataset collected is imperative for the efficiency of the model. The division of dataset into training and testing sets is also a deciding factor for the efficiency of the model. The training set is the main aspect upon which the results depends.

B. Pre-Processing of tweets

The pre-processing of the data is a very important step as it decides the efficiency of the other steps down in line. It involves syntactical correction of the tweets as desired. The steps involved should aim for making the data more machine readable in order to reduce ambiguity in feature extraction. Below are a few steps used for pre-processing of tweets –

1) Removal of Re-Tweets

a) Converting Upper Case to Lower Case: If we are using case sensitive analysis then we should take two occurrence of same words as different due to their sentence case. It is important for an effective analysis not to provide such misgivings to the model.

b) Stop Word Removal: Stop words that don't affect the meaning of the tweet are removed. WEKA machine learning package checks each word from the text against a dictionary.

c) Twitter Feature Removal: User names and URLs are not important from the perspective of future processing, hence their presence is futile. All usernames and URLs are converted to generic tags or removed.

d) Stemming: Replacing words with their roots, reducing different words with similar meanings which helps in reducing the dimensionality of the feature set.

e) Special Character and Digit Removal: Digits and special characters don't convey any sentiment. Sometimes they are mixed with words, hence their removal can help in associating two words which were otherwise considered different. Creating a dictionary to remove unwanted words and punctuation marks from the text. Expansion of slangs and abbreviations and spelling correction. Generating a dictionary for words that are important or for emoticons.

f) Part of Speech (POS) Tagging: It assigns tag to each word in text and classifies a word to a specific category like noun, verb, adjective etc. POS taggers are efficient for explicit feature extraction.

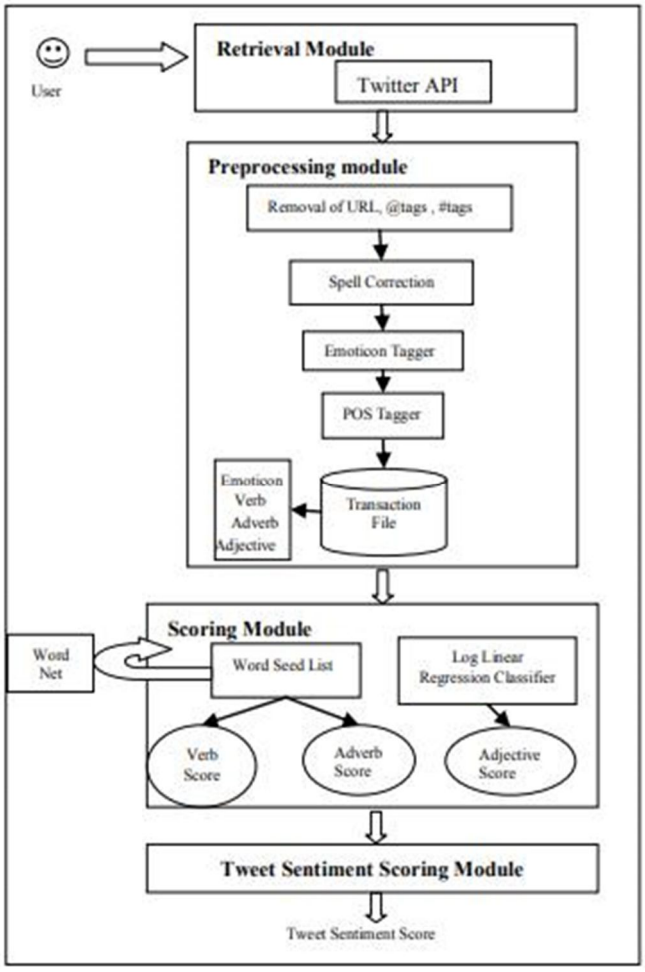
©IJRASET: All Rights are Reserved

C. Twitter Sentiment Analysis

The sentiment can be in the comments or tweets to provide useful indicators for many different purposes. The sentiments can be divided into two groups: negative positive

Sentiment analysis is a natural language processing techniques to quantify an expressed opinion or sentiment within a selection of tweets. 2014 Sentiment analysis refers to the general method to extract polarity and subjectivity from semantic orientation which refers to the strength of words and polarity text or phrases. There has two main approaches for extracting sentiment automatically which are the lexicon-based approach and machine-learning-based approach.

D. System Architecture



E. Machine-Learning-Based Approach

Machine learning methods often rely on supervised classification approaches where sentiment detection is framed as a binary which are positive and negative. This approach requires labelled data to train classifiers. This approach, it becomes apparent that aspects of the local context of a word need to be taken into account such as negative (e.g. not beautiful) and intensification (e.g. Very beautiful). However, showed a basic paradigm for create a feature vector is:

1) Apply a part of speech tagger to each tweet post.
2) Collect all the adjective for entire tweet posts.
3) Make a popular word set composed of the top N adjectives.
4) Navigate all of the tweets in the experimental set to create the following:
a) Number of positive words
b) Number of negative words

©IJRASET: All Rights are Reserved

Presence, absence or frequency of each word showed some example of switch negation, negation simply to reverse the polarity of the lexicon: changing beautiful (+3) into not beautiful (-3). More examples: She is not terrific (6-5=1) but not terrible (-6+5=-1) either. In this case, the negation of a strongly negative or positive value reflects a mixed perspective which is correctly captured in the shifted value. However, [21] has mentioned the limitation of machine-learning-based approach to be more suitable for Twitter than the lexical based method.

Furthermore, machine learning methods can generate a fixed number of the most regularly happening popular words which assigned an integer value on behalf of the frequency of the word in the Twitter.

F. Techniques of Sentiment Analysis

The semantic concepts of entities extracted from tweets can be used to measure the overall correlation of a group of entities with a given sentiment polarity. Polarity refers to the most basic form, which is if a text or sentence is positive or negative. However, sentiment analysis has techniques in assigning polarity such as:

1) *Natural Language Processing (NLP):* NLP techniques are based on machine learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules. Sentiment analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level. NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world.
2) *Artificial Neural Network (ANN):* It is known as neural network is a mathematical technique that interconnects group of artificial neurons. It will process information using the connections approach to computation. ANN is used in finding the relationship between input and output or to find patterns in data.
3) *Support Vector Machine(SVM):* Support Vector Machine is to detect the sentiments of tweets together with stated SVM is able to extract and analyse to obtain upto70%-81.3% of accuracy on the test set collected training data from three different Twitter sentiment detection websites which mainly use some pre-built sentiment lexicons to label each tweet as positive or negative. Using SVM trained from these noisy labelled data, they obtained 81.3% in sentiment classification accuracy.

G. Application Programming Interface (API)

Alchemy API performs better than the others in terms of the quality and the quantity of the extracted entities. As time passed the Python Twitter Application Programming Interface (API) is created by collected tweets. Python can automatically calculated frequency of messages being retweeted every 100 seconds, sorted the top 200 messages based on there-tweeting frequency, and stored them in the designated database . As the Python Twitter API only included Twitter messages for the most recent six days, collected the data needed to be stored in a different database.

H. Python

Python was found by Guido Van Rossum in Netherlands, 1989 which has been public in 1991. Python is a programming language that's available and solves a computer problem which is providing a simple way to write out a solution. Mentioned that Python can be called as a scripting language. Moreover, Python is a just description of language because it can be one written and run on many platforms. In addition, mentioned that Python is a language that is great for writing a prototype because Python is less time consuming and working prototype provided, contrast with other programming languages. Many researchers have been saying that Python is efficient, especially for a complex project, as has mentioned that Python is suitable to start up social networks or media steaming projects which most always are a web-based which is driving a big data because Python can handle and manage the memory used. Besides Python creates a generator that allows an iterative process of things, one item at a time and allow program to grab source data one item at a time to pass each through the full processing chain.

V. CONCLUSION

Twitter sentiment analysis is developed to analyse customers_ perspectives toward the critical to success in the marketplace. The work presented in this paper specifies a machine learning based approach with natural language processing techniques for sentiment analysis on Twitter data. To associate with Twitter API, developer need to agree in terms and conditions of development Twitter platform which has been provided to get an authorization to access a data. The output from this process will be saved in JSON file. The reason is, JSON (JavaScript Object Notation) is a lightweight data-interchange format which is easy for humans to write and

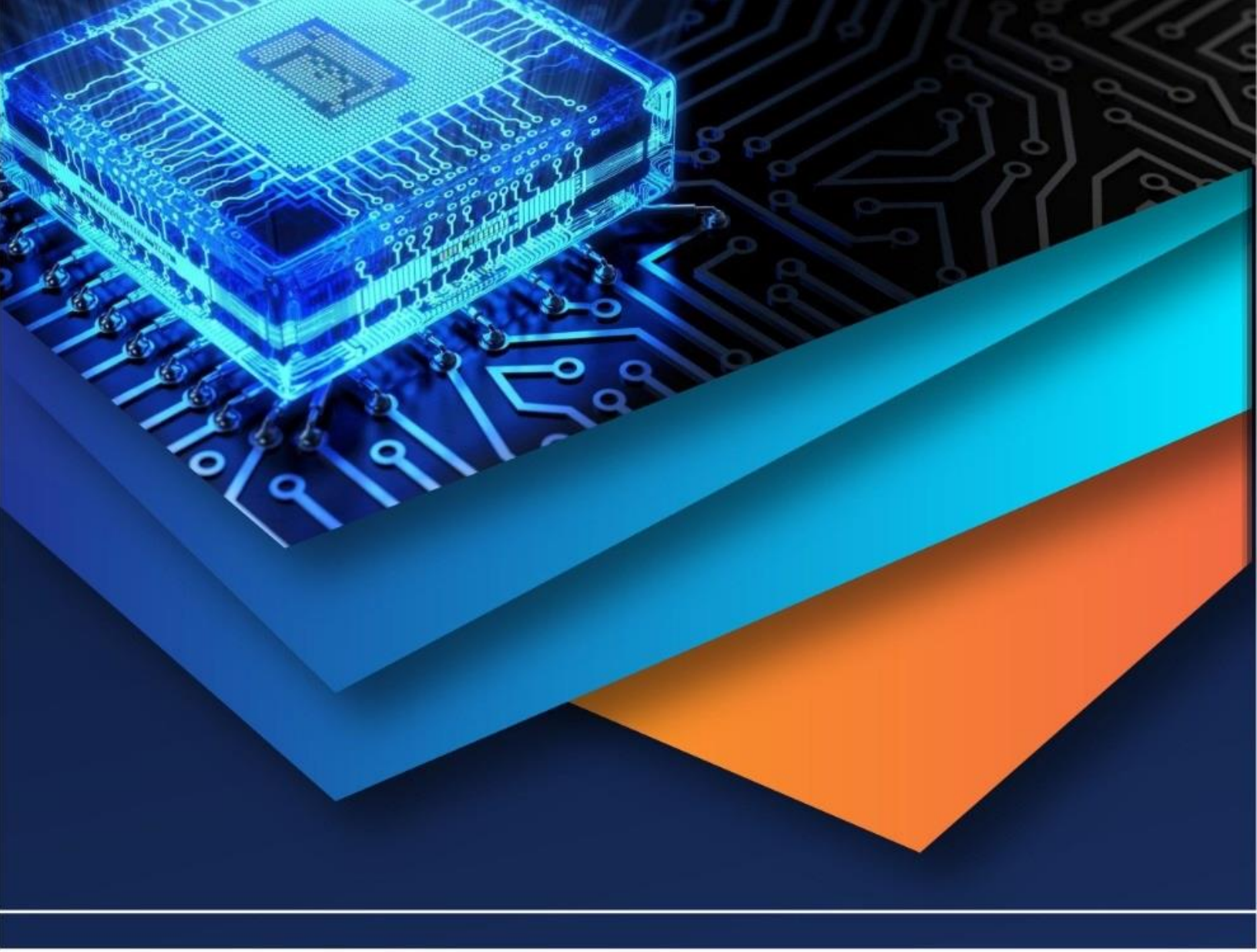
©IJRASET: All Rights are Reserved

read. Tweets from JSON file will be assigned the value of each word by matching with the lexicon dictionary. As a limitation of words in the lexicon dictionary which is not able to assign a value to every single word from tweets. However, as a scientific language of python, which is able to analyse a sense of each tweet into positive or negative for getting a result.

REFERENCES

[1] Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python", International Journal of Computer Applications (0975 – 8887).
[2] Aliza Sarlan, Chayanit Nadam, Shuib Basri, "Twitter Sentiment Analysis", 2014 International Conference on Information Technology andMultimedia (ICIMU), November 18 – 20, 2014.
[3] Pierre FICAMOS*, Yan LIU, "A Topic based Approach for Sentiment Analysis on Twitter Data", (IJACSA) International Journal of Advanced Computer Science and Applications.
[4] Akshi Kumar and Teeja Mary Sebastian, "Sentiment Analysis on Twitter", IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814.
[5] Hamid Bagheri and Md Johirul Islam, "Sentiment analysis of twitter data".
[6] Onam Bharti and Mrs. Monika Malhotra, "Sentiment Analysis on Twitter Data", International Journal of Computer Science and Mobile Computing (IJCSMC), June-2016.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)