iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 4 **Issue:** V **Month of publication:** May 2016

**DOI:**

www.ijraset.com

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

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# Sentiment Analysis on Twitter by using Machine Learning Technique

Neha Upadhyay[1], Prof. Angad Singh[2]
*Department of Information and Technology, Bhopal*

***Abstract:*** ***Due to the vast opinion of rich web resources such as discussion forum, review sites , blogs and news corpora on the market in digital form, a lot of analysis is focusing on the area of sentiment analysis. People are trying to develop a system that can identify and classify opinion or sentiment as represented in an electronic text. A correct method to predict sentiments could modify us, to extract opinions from the internet and predict online customer's preferences, might prove valuable for economic and market research. Till now, there are few different issues predominating in this analysis, i.e, sentiment classification, feature based classification and handling negations. In this thesis, we try to estimate the twitter posts about electronic products like mobiles, laptops etc.we collect the tweets of mobile from twitter and preprocess the tweets, in preprocessing remove irrelevent words such as name, symbols etc. after preprocessing we catogarized the tweets in two parts opinion or opinionless, we compare each tweets to database of positive, negative and average words those tweets carry positive, negative and average words, is in opinion cateogity other's discarted. now we check the polarity of opinion tweets, count the polarity of tweets and divide in positive, negative and average category. Estimating the probablity of positive tweets we use Naive bayes and S.V.M and maximum entropy and compare results for better probablity result.***

***Keywords: Tweets,Sentiment Analysis, Machine Learning Techniques, Naive Bayes, SVM, Maximum Entropy, Ensemble Classifier.***

## I. INTRODUCTION

The period of Internet has changed the way individuals express their perspectives. It is currently done through blog entries, online disscusion forums, item review sites etc. people rely on this client created data as it were. When people needs to purchase an item, they will lookup its surveys online before taking a choice. The measure of client created content is too large for an ordinary client to analyse. So to solve this, different sentiment analysis have used[1]. Symbolic methods or Knowledge base approach and Machine learning strategies are the two primary procedures are use as a part of opinion analysis. Knowledge base approach requires an expansive database of predefined emotions and a efficient learning representation for identifying sentiments[2]. Machine learning approach makes use of a training set to build up a sentiment classifier that classify opinions. Since a predefined database of whole feelings is not required for machine learning approach, it is more simpler than Knowledge base methodology. In this paper, We used different machine learning procedures for classifyng tweets. Sentiment analysis is normally used at various levels fluctuating from coarse level to fine level.[3] Coarse level assumption examination manages deciding the estimation of a whole archive and Fine level manages quality level slant investigation. Sentence level estimation investigation comes in the middle of these two. There are numerous looks into on the territory of opinion investigation of client audits. Past examines demonstrate that the exhibitions of slant classifiers are subject to points. Due to that we can't say that one classifier is the best for all subjects since one classifier doesnt reliably outflanks the other. Conclusion Analysis in twitter is very troublesome because of its short length. Nearness of emoticons, slang words and incorrect spellings in tweets compelled to have a preprocessing venture before highlight extraction. There are diverse component extraction strategies for gathering pertinent elements from content which can be connected to tweets moreover. In any case, the element extraction is to be done in two stages to extricate applicable elements. In the primary stage, twitter particular components are removed. At that point these elements are expelled from the tweets to make ordinary content. After that, again include extraction is done to get more components. This is the ticket utilized as a part of this paper to create an effective component vector for investigating twitter assessment.[4][5] Since no standard dataset is accessible for twitter posts of electronic gadgets, we made a dataset by gathering tweets for a specific timeframe.

By doing assessment investigation on a particular area, it is conceivable to distinguish the impact of space data in picking an element vector. Distinctive classifiers are utilized to do the grouping to discover their impact in this specific space with this specific element vector.



©IJRASET : All Rights are Reserved

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

## II. RELATED WORK

There are two fundamental approaches to identify sentiment from texts. They are Symbolic methods and Machine Learning methods. The following two sections manage these techniques.

### *A. Symbolic Techniques*

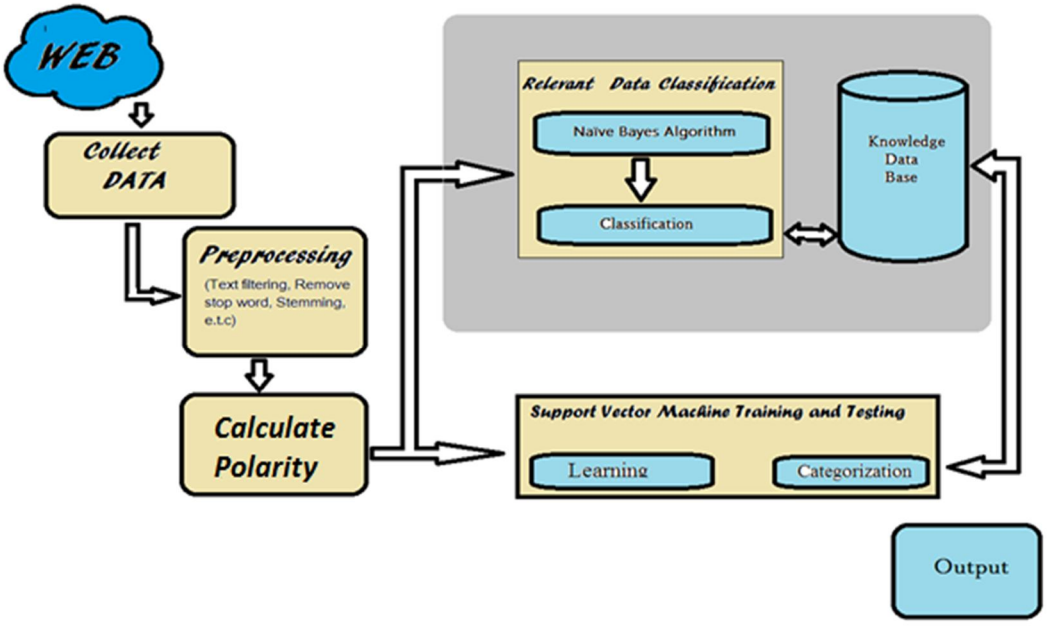
A significant part of this research in unsupervised sentiments classification by using symbolic methods makes use of accessible lexical resources. Turney used pack of-words methodology for sentiment analysis. In this approach, connections between the individual words are not considered and a report is represented as a insignificant collection of words. To determine the general sentiment, opinion of every word is resolved and those values are combine with some aggregation functions. [6]He found the polarity of a survey in view of the normal semantic introduction of tuples separated from the survey where tuples are expressions having descriptive words having adverbs or adjectives. He found the semantic introduction of tuples using the web search engine Altavista

### *B. Machine Learning Techniques*

Machine Learning strategies used a training set and a test set for a classification. Training set contains information feature vectors and their corresponding class labels. By using this training set, a classification model is created which tries to classify the information feature vectors into corresponding class labels.[7] At that point a test set is used to accept the model by predicting the class labels of unseen feature vectors. Various machine learning merthods like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews. Some features that can be used for opinion classification are Term Presence, Term Frequency, negations, n-grams and Part-of-Speech[8][9] . These features can be used to identify the semantic introduction of words, expressions, sentences and that of documents. Semantic introduction is the polarity which may be either positive or negative.

## III. PROPOSED ARCHITECTURE

The proposed design of four modules: user interface, log preprocessing, Naïve Bayes Classification,Training and testing use support vector machine for more accuracy of opinion categorization . This framework can solve irrelevant data and more accurateness by associating svm with the Naïve Bayes Classification.[10][11]





©IJRASET : All Rights are Reserved

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

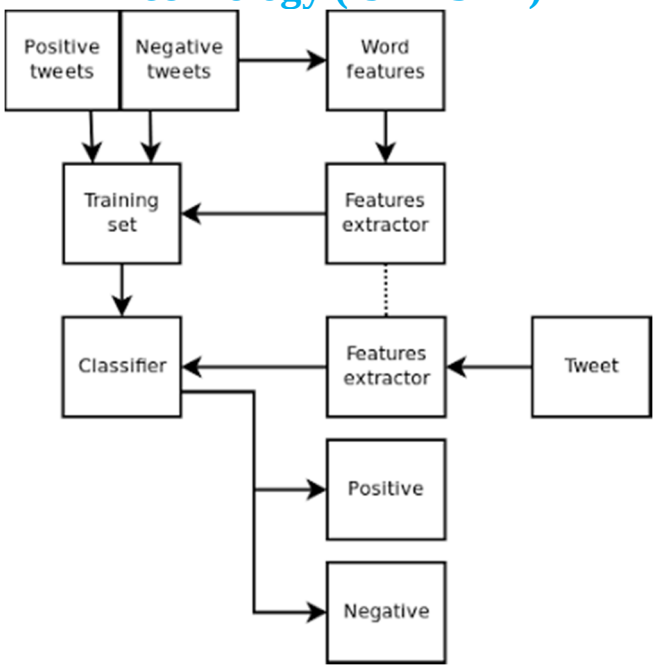



Fig1 and 2: Proposed Architecture

*A. Text Pre-Processing*

Text Pre-preparing for extract significant data.

1) *Stop words* - an, is, the, with and so on. The full list of stop words can be originate at Stop Word List. These words don't show any sentiments and can be removed.
2) *Repeating letters* - in the event that you take a look at the tweets, in some cases peaple repeat letters to stretch the feeling. E.g. hunggrryyy, huuuuuuungry for 'hungry'. We can search for 2 or more repeative letters in words and replace them by 2 of the same.
3) *Punctuation* - we can remove punctuations like comma, single/double quote, question marks toward the begin and end of every word. E.g. excellent!!!!!! replaced with excellent

## IV. METHODOLOGY USED

*A. Nave Bayes Classifier*

The Naive Bayesian classifier depends on Bayes' theorem with independent assumptions between indicators. A Naive Bayesian model is easy to work , with no confused iterative parameter estimation which makes it especially valuable for large datasets. The conditional probability for Naive Bayes can be characterized as

$$P(X|y_j) = \Pi_{i=1}^{m} P(x_i|y_j)$$

"X" is the feature vector characterized as X=fx1,x2,....xmg and yj is the class label. Here, in our work there are various idependent features like emoticons, emotional keywords, count of positive and negative keywords, and check of positive and negative hash labels which are successfully used by Naive Bayes classifier for classification. Nave Bayes does not consider the connections between components. So it can't use the connections between part of speech tag, emotional keyword and negation.

*B. SVM Classifier:*

A Support Vector Machine (SVM) is a discriminative classifier formally characterized by an seperating hyperplane.A discriminative function characterized as


©IJRASET : All Rights are Reserved

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

$$g(X) = w^T \phi(X) + b$$

"X" is the feature vector, "w" is the weights vector and "b" is the bias vector. _() is the non linear mapping from input space to high dimensional feature vector. "w" and "b" are found automatically on the training set. Here we used a linear kernel for classification. It maintains a wide gap between two classes.

Support Vector Machines is a well known classifying method. We use the SVM lightsoftware with a linear kernel. Our input data are two sets of vectors of size m. Every section in the vector corresponds to the presence of a feature. For example, with a unigram feature extractor,each feature is a single word found in a tweet. If the feature is present, the value is 1, however the feature is missing, then the value is 0. We use feature presence, instead of a count, so that we do not have to scale the input data, which speeds up overall processing.

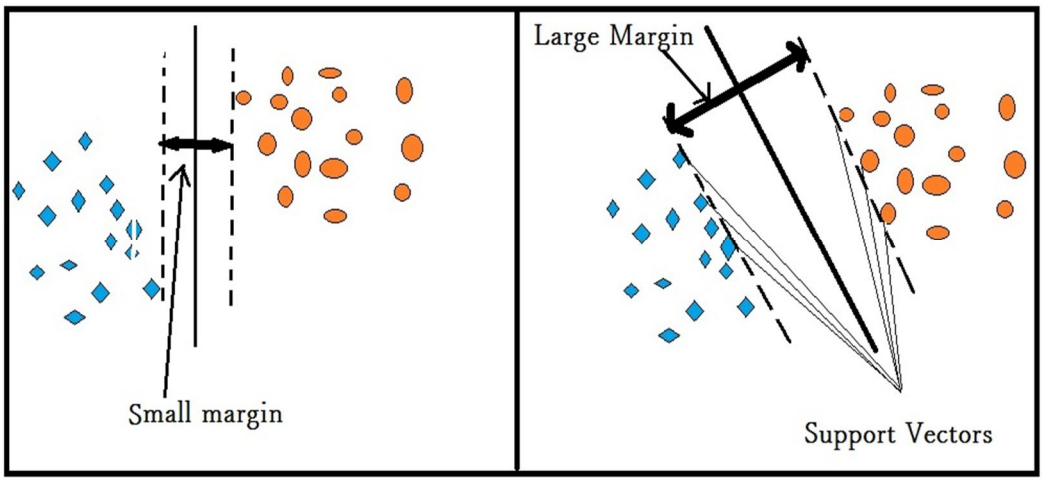



**Fig3. S.v.m.**

*C. Maximum Entropy Classifier*

In Maximum Entropy Classifier, no assumptions are taken regarding to the relationship between features. This classifier always try to maximize the entropy of the system by assessing the conditional circulation of the class label. The conditional distribution is characterized as

$$P_\lambda(y|X) = 1/Z(X) exp\left\{\sum_i \lambda_i f_i(X,y)\right\}$$

'X' is the feature vector and 'y' is the class label. Z(X) is the normalization factor and _i is the weight coefficient. fi(X,y) is the feature function which is defined as

$$f_i(X,y) = \begin{cases} 1, & \text{X=x}_i \text{ and } y = y_i \\ 0, & \text{otherwise} \end{cases}$$

In our feature vector, the relationships between part of speech tag, emotional keyword and negation are utilized effectively for classification.

*D. Ensemble classifier*

Ensemble classifiers can be of various types. They try to make use of the features of all the base classifiers to do the best classification. The base classifiers used here are Nave Bayes, Maximum entropy and SVM. Here an ensenble classifier is created by voting principle. The classifier will classify based on the output of majority of classifiers.



©IJRASET : All Rights are Reserved

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

## V. RESULT AND ANALYSIS

This part examines and evaluates the observational results of tests to validate the exhibited framework proposed. The reasonable motive of the work is accomplished in this section

Firstly, choose the Product for which we require to analyse the sentiment and after that calculate the Positive Probability of Product survey through any of them classifiers.

In this part we figure out the importance of the proposed approach. In proposed approach we are using two classifiers: Naïve Bayes and S.V.M

Almost 2000 tweets were classified into three different classes: Positive Sentiment, Negative Sentiment and Average Sentiment. Figure 4 shows classification of Product tweets into three classes (positive, negative and Average) the help of segment charts. We have taken Electronic Product tweets for our research work. These 6 Products are the best product of year 2014 – 2015.

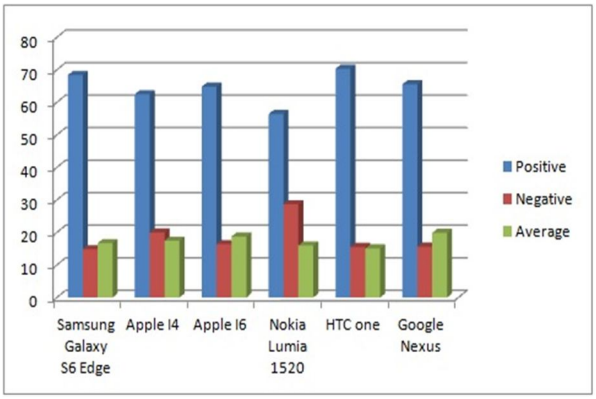



Fig4: Classification of individual product tweets

And Fig 4,5,6 shows the classification of 6 phones into positive and negative sentiments.

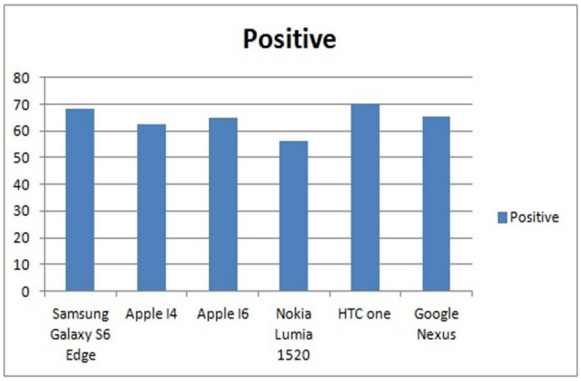



Figure5: Classification of individual Product Positive tweets

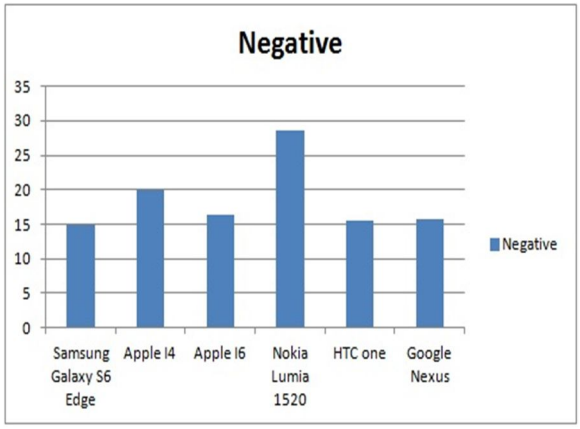



Figure6: Classification of individual Product Negative tweets



©IJRASET : All Rights are Reserved

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

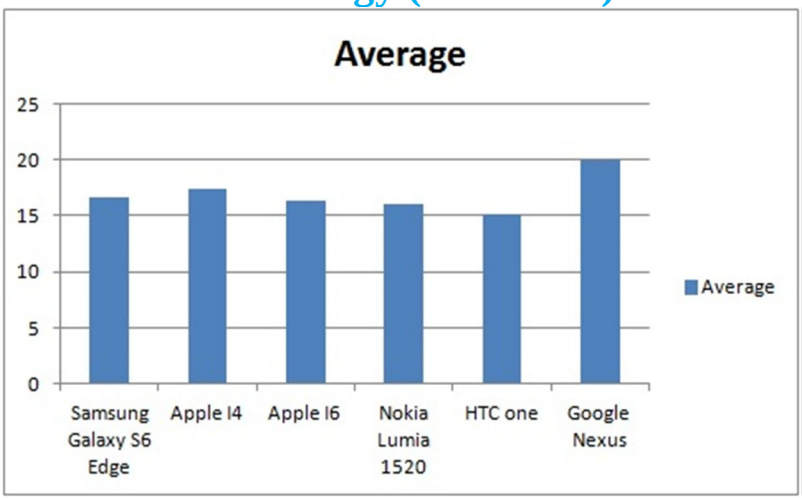



Figure7: Classification of individual Product Average tweets

We can figure our outcome through Naïve Bayes and S.V.M, and compare the outcome with the help of positive sentiments, negative sentiments and neutral sentiments. We can compute our outcome through Naïve Bayes and S.V.M classifiers and compare the result with the help of positive sentiments, negative sentiments and neutral sentiments.
This line diagram explain the comparision of Naïve Bayes And S.V.M classifier for positive tweets of three Product (Samsung Glaxy S6, Apple I4, Apple I6).

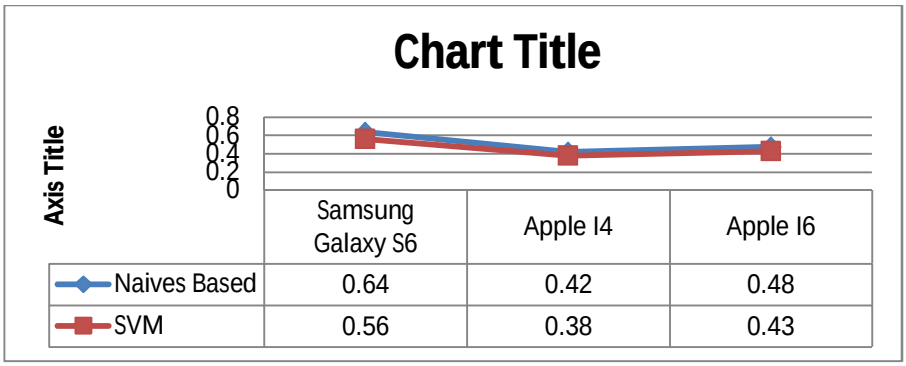



Fig.8 Comparison between Naive Bayes and S.V.M on Positive review

This line diagram clarifies the comparision of Naïve Bayes and S.V.M classifier for neutral tweets of three product.The above graphs are very helpful for representing the result of proposed system. Table in fig 9 shows the sentiments of all product tweets in a table after filtering the tweets.

| | Positive % | Negative % | Average % |
|---|---|---|---|
| Galaxy S6edge | 68.41 | 14.86 | 16.72 |
| Apple I4 | 62.53 | 20.00 | 17.46 |
| Apple I6 | 64.82 | 16.40 | 18.77 |
| Nokia Lumia 1520 | 56.42 | 28.68 | 16.00 |
| HTC One | 70.34 | 15.56 | 15.10 |
| Google Nexus | 65.58 | 15.62 | 15.10 |

Fig.9 Table of Sentiments of tweets after polarity calculation

## VI. CONCLUSION

There are different procedures to identify the sentiments from content. In this paper,our analysis represents that Machine Learning


©IJRASET : All Rights are Reserved

strategies are less difficult and more efficient. This technique can be applied for twitter sentiment analysis. Theresome issues while dealing with identifying setiment keyword from tweets having multiple keywords. It is also challanging to solve missspelt and slang words.To manage this issue,to solve the difficulties in data, an efficient feature vector is made by doing feature extraction in two stages after appropriate preprocessing. Classification accuracy of the feature vector is tested by using different classifiers like Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers.This feature vector performs useful for electronic products.The message communicated in Twitter can be identified with the human behaviour, nature, personality and attitude. Classification of tweets into positive sentiments, negative sentiments or neutral sentiments indicates the views of people on Product. That's help people to choose best item and they easily decide that which item is famous in the overall business market.

## REFERENCES

[1] H. Isah, P. Trundle and D. Neagu, "Social networking media identifying for product safety using data mining and sentiment analysis," *Computational Intelligence (UKCI), 2014 14th UK Workshop on*, Bradford, 2014, pp. 1-7.

[2] M. Anjaria and R. M. R. Guddeti"Influence factor based opinion mining of Twitter using supervised learning process," *(COMSNETS) Communication Systems and Networks, 2014 Sixth International Conference on*, Bangalore, 2014, pp. 1-8.

[3] Tiara, M. K. Sabariah and V. Effendy, "Sentiment analysis on Twitter using lexicon-based and support vector machine methodology for assessing the performance of a television program," *( ICoICT ) Information and Communication Technology, 3rd International Conference on 2015*, Nusa Dua, 2015, pp. 386-390.

[4] R.Rajshree and M S. Neethu," machine learning techniques used in sentiment analysis in twitter," *(ICCCNT)Computing, Communications and Networking Technologies, Fourth International Conference 2013 on*, Tiruchengode, 2013, pp. 1-5.

[5] G. Gautam and D. Yadav, ."Sentiment analysis in twitter using semantic analysis and machine learningapproaches," *Contemporary Computing (IC3), 2014 Seventh International Conference on*, Noida, 2014, pp. 437-442.

[6] J. Akaichi, Z. Dhouioui and M. J. Lopez-Huertas Perez, "For sentiment classification text mining face book updates on," *(ICSTCC) System Theory, Control and Computing, 2013 17th International Conference*, Sinaia, 2013, pp. 640-645.

[7] M. S. Schlichtkrull, "Learning affective projections for emoticons on Twitter,*6th IEEE International Conference*," *(CogInfoCom) Cognitive Infocommunicationson*, Gyor, 2015, pp. 539-543.

[8] D. Hakkani-Tür and A. Celikyilmaz, J. Feng, "Probabilistic model-based sentiment classification of twitter texts," *(SLTTechnology of Spoken Language Workshop, IEEE 2010*, Berkeley, CA, 2010, pp. 79-84.

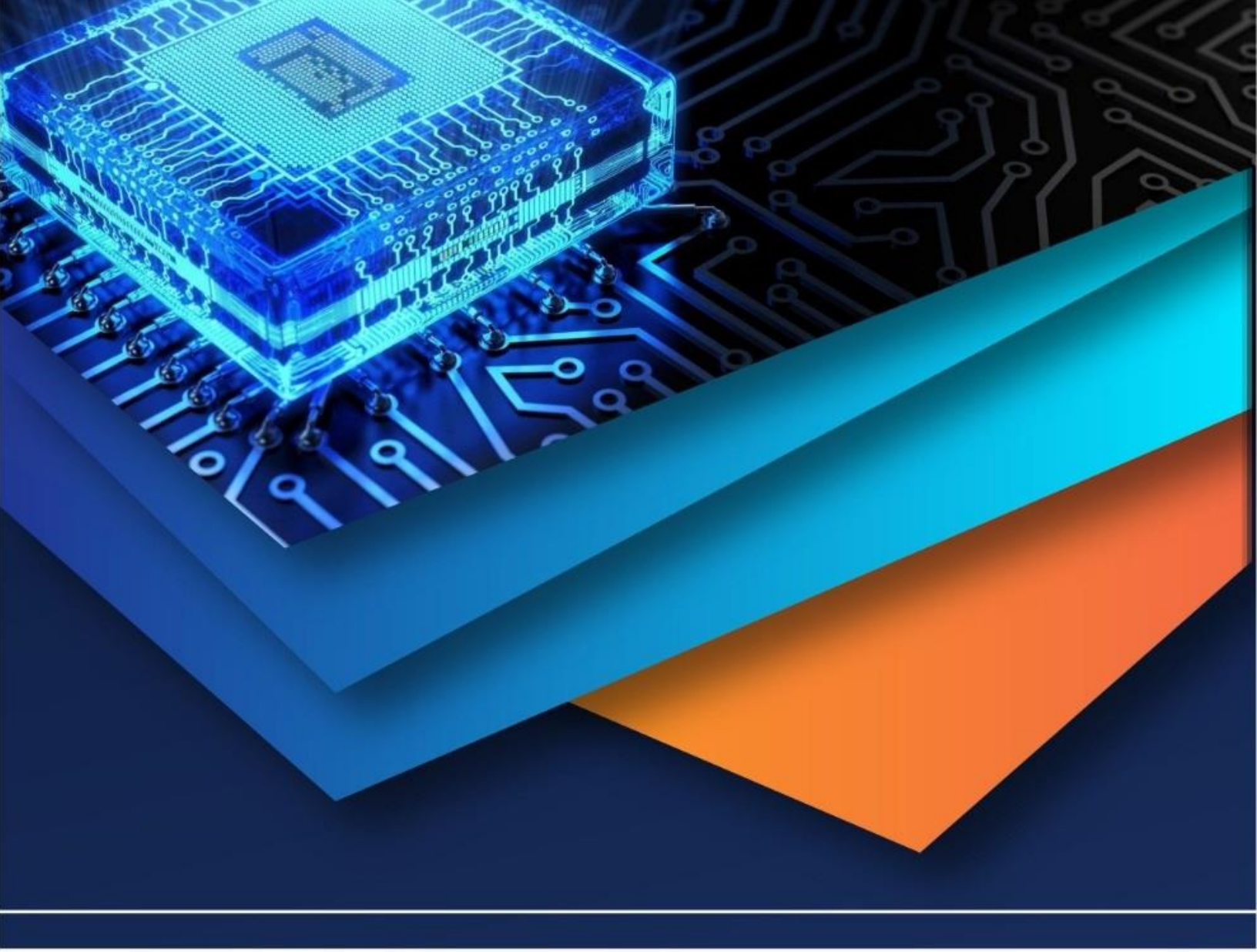
[9] Q. H. Vuong and A. Takasu," Transfer Learning for Emotional PolarityClassification," *(IAT) Intelligent Agent Technologies and (WI) Web Intelligence , 2014 IEEE/WIC/ACM International Joint Conferences on*, Warsaw, 2014, pp. 94-101.

[10] N. Azam, Jahiruddin, M. Abulaish and N. A. H. Haldar ,"Twitter Data Mining of Events Analysis and Classification," *2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI)*, Hong Kong, 2015, pp. 79-83.

[11] M. Kanakaraj and R. M. R. Guddeti," *Semantic Computing (ICSC), 2015 IEEE International Conference on*, Anaheim, CA, 2015, pp. 169-170.



©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)