iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

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 :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Comparative Study for Sentiment Analysis of Twitter Data

Rishija Singh[1], Dr. Vikas Goel[2]
[1, 2]Ajay Kumar Garg Engineering College

Abstract: ***Sentimental Analysis is reference to the task of Natural Language Processing to determine whether a text contains subjective information and what information it expresses i.e., whether the attitude behind the text is positive, negative or neutral. Social media platforms and micro blogging websites are the rich sources of user generated data. Through these resources, users from all over the world express and share their opinions about a variety of subjects. The analysis of such a huge amount of user generated data manually is impossible, therefore an effective and intelligent technique is needed which can analyze and provide the polarity of this textual data. Multiple tools and techniques are available today for automatic sentiment classification for this user generated data. Mostly, three approaches are used for this purpose Lexicon based techniques; Machine Learning based techniques and hybrid techniques (which combines lexicon based and machine learning based approach).The purpose of this study is to explore the different machine learning techniques to identify its importance as well as to raise an interest for this research area.***

Keywords: Social networking, micro-blogging, Twitter, sentiment analysis

I. INTRODUCTION

The age of Internet has changed the way people express their views. It is now done through blog posts, online discussion forums, product review websites etc. People depend upon this user generated content to a great extent. When someone wants to buy a product, they will look up its reviews online before taking a decision. The amount of user generated content is too large for a normal user to analyze. So to automate this, various sentiment analysis techniques are used. Sentimental Analysis is interpreted as determining the notion of people about distinct existence. Nowadays people are used to review the comments and posts on the product which are known as opinion, emotion, feeling, attitude, thoughts or behavior of the user. Sentimental Analysis is a method for identifying the ways in which sentiment is expressed in texts. Sentimental analysis attempts to divine the posture or notion of a keynoter or author, or author against assertive field or an object. There are many claims in sentiment analysis. First is that, a viewpoint which is treated as positive in one case and will be taken as negative in another case. The next claim is that usually people don't consider their viewpoint in same form. Almost of all reviews incorporate with both positive as well as negative remarks, which can be feasible by interpreting the sentences each at a time. Finding the opinion sites and monitoring them on the web is somewhat difficult. So there will be a need of robotic opinion mining as well as a Summarization system. In sentiment analysis there are three classification levels: document-level classification, sentence-level classification and feature-level sentiment analysis. In document-level classification the main intention is to classify an opinion in the whole document as positive and negative. It speculates entire document as a single unit. The aim of sentence-level analysis is to categorize emotion expressed in respective sentences. In sentence-level the basic step is to recognize the sentence as objective or subjective. Suppose sentence is subjective, it will decide whether it express negative or a positive opinion. In aspect-level analysis it aims to categorize the sentiment in respect of particular entities. Generally, there are two approaches in sentimental analysis. One is by considering symbolic methods and other one by machine learning method. In symbolic learning technique, this is categorized according to some learning strategies such as learning from analogy, discovery, and examples and from root learning. In machine learning technique it uses unsupervised learning, weakly supervised learning and supervised learning. Along with lexicon based and linguistic method, machine learning will be considered as one of the mainly used approach in sentiment classification.

A. Sentiment Analysis and Opinion Mining

The study of people's point of view or emotions towards a product or an event is "Sentiment Analysis". Sentiment analysis helps to track the reputation of product or services in general. Sentiment classification can be at sentence level or document level. Document

level classification needs to filter out the sentences that don't contain opinion words before classifying it into positive or negative. The method for classifying the phrases first extracts the opinionated text, then estimates the positions of these texts in the phrases and finally positive or negative value is assigned to the given phrase.

B. Twitter

There are almost 111 micro-blogging sites today over the internet. These micro-blogs are actually social media that that the people use to share their posts. Among the 111 micro-blogs, twitter is one of the most popular sites. Twitter lets the people post tweets (message) of 148 characters in length Micro-blogging websites are social media that helps users to make short and frequent posts. As one tweet only consists of 148 characters, it makes the process of sentiment analysis easier.

II. SENTIMENT CLASSIFICATION TECHNIQUES

The main twitter sentiment classification techniques are Support Vector Machines (SVMs), Naïve Bayes Classifier, Fuzzy logic, Baseline Model, Maximum Entropy, K-Nearest Neighbor Classifier, Baseline model.

A. Naive Bayes Classifier

A Naive Bayesian classifier is one of the familiar supervised learning techniques which are frequently used for classification purpose. Their classifier is named as naïve since it considers the contingency that are actually linked are not depending on the further. Calculation of whole document feasibility would be the substance in aggregation of all the feasibility report of single word in the file. These Naïve Bayesian classifiers were frequently applied in sentiment categorization since they are having lower computing power when comparing to the other approach but independence assumptions will provide inaccurate results.

B. Support Vector Machine

Support Vector Machine (SVM) is known as the best classifier that provides the most accurate results in speech classification problems. They achieved by creating a hyper plane with maximal Euclidean distance for the nearest trained examples. Support Vector Machine hyper plane are completely resolved by a comparatively minute subset of the trained data sets which are treated as support vectors. The remaining training data sets have no access on the qualified classifier. So for the purpose of text classification, the classifier SVMs have been applied successfully and also used in different sequence processing application. SVMs are used in hypertext and text classification since they do not require labeled training data set.

C. Bayesian Network

The main disadvantage of Naïve Bayes classifier is its independent assumption of aspects in data sets. This assumption is the reason for start of using Bayesian Network. This Bayesian network is directed non-cyclic graph where nodes correspond to variables and those edges are correspond to conditional independency. In text classification Bayesian
Network is not usually used since it is expensive in computation.

D. Maximum Entropy

Maximum Entropy classifier is parameterized by a weight set that are used to associate with the joint-future, accomplished by a trained data set by encoding it. This Maximum Entropy classifier appear with the group of classifiers such as log-linear and exponential classifier, as its job is done by deriving some data sets against the input binding them directly and the result will be treated as its exponent.

E. K- Nearest Neighbor Classifier

K-Nearest Neighbor is a unsupervised learning algorithm for text classification. In this algorithm the entity is classified with various trained data set along with their nearest distance against each entity. The advantage with this algorithm is its simplicity in text categorization. It also works well with multi-class text classification. The main drawback of KNN is it necessitate with large amount of time for categorizing entities where huge data set are inclined.

F. Fuzzy Logic

Fuzzy logic is used to draw and retrieve sentiments from a text or document. Fuzzy logic uses the concept of reasoning that gives results in approximation rather than exact results. Fuzzy logic is useful for managing such approximate information. The numerical score of sentence is evaluated between the range from 0 to 1.

G. Baseline Model

In baseline model, initially the preprocessing steps are carried out study the polarity frequencies of unigrams, bigrams, and trigrams in the training data set. Three probability score is given to each token: Neutral probability, Positive probability and negative probability. A feature vector is then created for the tokens that can differentiate the tweet's sentiment effectively. Before the probability values are calculated the infrequent words are filtered out, this serves the baseline model. For example, Emotion Determiners present with value 1 indicates its presence in the text and 0 indicates its absence in the text. Various features are appended to this model after building it.

H. Corpus and Dictionary Based

In dictionary based approach, first the seed words of opinions are searched and then it looks for their synonyms and antonyms. Some of the opinion words are listed manually. The list is then expanded by searching into popular or well-known corpora like WordNet. The strength of polarity is also listed in the dictionary for each word. The corpus based approach is use to find opinion words with context-specific orientations which depends on syntactic patterns.

III. DIFFERENT RESEARCHES ON SENTIMENT ANALYSIS

A. Sentiment Analysis on Twitter Data

V. Sahayak, V. Shete and A. Pathan [3] proposed in 2015 about "Sentiment Analysis on Twitter Data" suggested the hybrid approach that classifies the tweets from twitter dataset in sentiment categories like positive , negative, and neutral. The two techniques used in this approach are corpus based, dictionary based sentiment classification, and it includes POS for polarity features and tree kernel to avoid monotonous features. The feature extractor & different machine learning classifier are explained and used in this methodology. The machine learning classifiers are Naïve Bayes, Support Vector Machines (SVM), and Maximum Entropy. These classifiers are used to make three models for feature extraction process namely unigram model, tree kernel model, feature based model. They developed the process of sentiment analysis of tweets, which contains three sections. First section is data extraction, which helps to extract opinion words from tweets. Second section does preprocessing of all the extracted words, which includes emoticons handling, filtration, tokenization, removal of stop words, n-gram construction. Third section classifies the sentiments using machine learning classifiers. This section works in two steps: 1. Model construction. 2. Model usage to check accuracy of classification. When complexity of emoticons and opinions increases, it becomes difficult for this approach to give right answer. This can be a drawback. For example, "The product was awesome but the services were gruesome". In this case, this approach may get confused for the result of sentiment.

B. Sentiment Analysis using Fuzzy Logic

Md. A. Haque&T. Rahman[1] proposed in 2014 "Sentiment Analysis with the help of Fuzzy Logic" by ranking the review in terms of positive and negative is the ranking perspective and it is achieved using fuzzy logic. The need of sentiment analysis is based on the two sectors i.e. classification of documents according to the orientation of sentiments such as positive and negative, other sector is gathering information by identifying the subjective or objective (SO) polarity of the comment or post, identifying the positivity or negativity (PO) polarity of comment or post and by identifying the degree of PN-polarity in terms of good, better or best. The tool to determine the polarity of lexical (the sentence is converted into sequence of tokens) is Senti Wor dNet. This gives numerical score to token range from 0 to 1. By having the values of the post and the weights ,the result can be computed by calculating the weighted and arithmetic mean, from that percentage of the individual sentiment(subjective &objective) is declared and by using concept of normalization the results are present in better way.

C. Sentiment Analysis on Twitter

A.Kumar and T. M. Sebastian[5] proposed in 2012 about "Sentiment Analysis on Twitter" developed hybrid approach using dictionary based and corpus based method to calculate semantic score of the opinion words in tweets. This approach uses the features like capitalization, emoticons, etc. while preprocessing the tweets. The approach more focuses on opinion words which must be a combination of adjectives and verbs. In this hybrid approach, dictionary based method is used to find semantic score of adverb, and verb. And corpus based method is used to find sematic score of adjectives. The list of adverbs and verbs along with their sematic strengths ranging from -1 to1 are taken into consideration while calculating semantic score of adverbs and verbs. The varying semantic strenghts of words provides high accuracy while handling multiple opinions and emoticons. For example, "very good" will get more semantic strength than "good" . The negation handling is also achieved by this approach and it is accurate. The

linear equation which is the highly focused part of this approach is used to calculate overall semantic score of single tweet. This linear equation calculates semantic score of each tweet more accurately by considering uppercase tweets, repeated letters, exclamation marks, emoticons, adjective group, verb group. According to semantic score of tweet, tweet is classified into three categories namely positive, negative, neutral.

D. A Fuzzy Logic Based on Sentiment Classification

J.I.Sheeba and Dr.K.Vivekanandan[5] proposed in 2014 "A Fuzzy Logic Based on Sentiment Classification" which says that, fuzzy logic is a type of probabilistic logic and it deals with reasoning that is approximate rather than fixed. Fuzzy logic is used for dealing with heterogeneous or vague information. Traditional logic may have many values but fuzzy logic can have values that range from 0 to 1. The input to the fuzzy logic is from sentiment classification step. The weights are assigned for each word. Based on the weight of the word, the "Threshold" value is calculated. The threshold value is calculated based on the average of each listed word. Finally, the list of positive, negative and neutral words are listed which is greater than or equal to the Threshold value. The algorithm is Fuzzy C-means algorithm use in this method which gather all the same words to reduce the emotions list and group the emotions based on the cluster centroid. Fuzzy approaches have started to emerge for text processing. In 2012, a review of fuzzy approaches for natural language processing highlighted that the percentage of papers relating to fuzzy approaches is very low over all the papers in the literature of natural language processing despite the suitability of fuzzy approaches for textprocessing and classification.

E. Sentiment Analysis and Opinion Mining

Y. Sharma, V. Mangat and M. Kaur [4] proposed in 2015 about "Sentiment Analysis & Opinion Mining" that suggested various approaches based on which the sentiments can be analyzed.

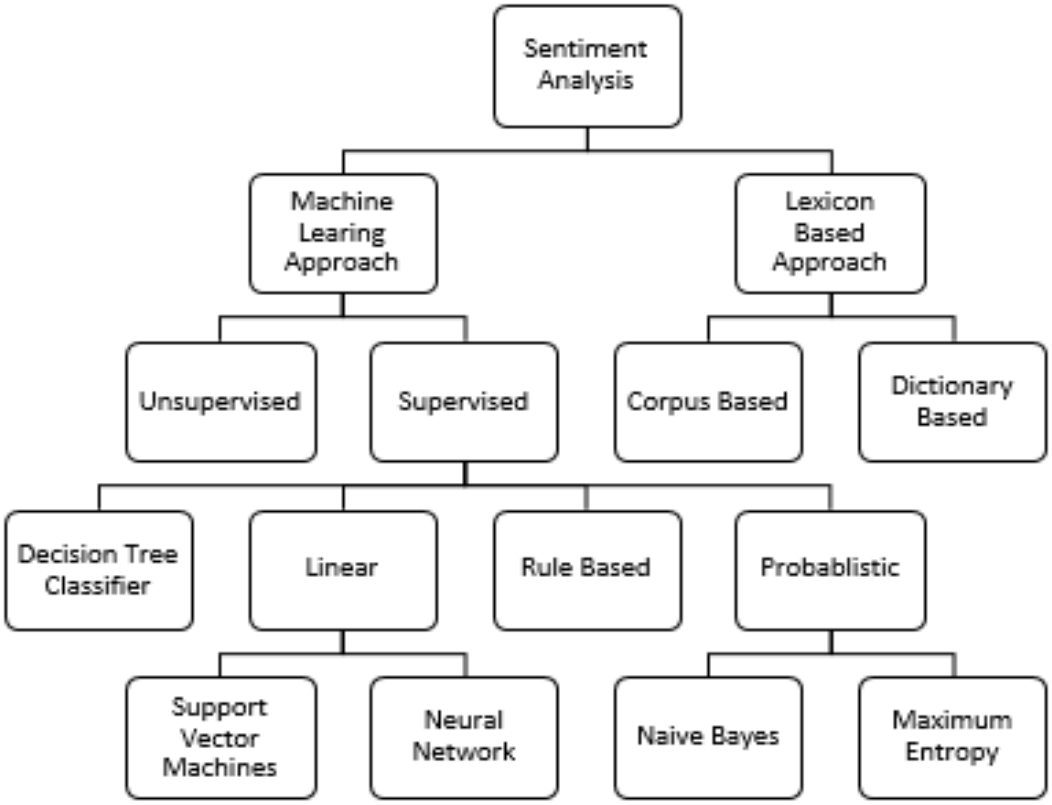


Fig. 1: Sentiment Classification Technique

The different approaches for sentiment analysis are shown in figure 1. It also includes the feature selection method, which reduces the unconnected information. It enhances the classification accuracy, and also decreases the running time of algorithm. The selection step is to remove the target, stop words, URL, & stemming. The two broad ways in which the sentiment analysis is done are Machine learning approach and Lexicon based approach. Machine learning approach learns from the previously generated results whereas the Lexicon based approach is usually fixed and gives approximated results.

F. A Hybrid Approach for Twitter Sentiment Analysis

N. Mittal & B. Agarwal [7] proposed "A Hybrid Approach for Twitter Sentiment Analysis" which is a three stage hierarchical model for sentiment extraction, in first stage the emoticons are labeled, then tweets are assigned sentiments using pre-defined lists of words with polarity and finally based on subjectivity of lexicon, the proposed probability based method assign the weight to all the tokens. The lexicons are weighted using various approaches like SentiWordNet, proposed probability based method, SentiWordNet (SWN) then probability based method, probability based method then SentiWordNet or Hybrid approach. The accuracy measured for hybrid approach was comparative higher than all the approaches i.e. 72.563 %. Hybrid method uses both SWN and probability based method to calculate the polarity of the token. Hence proposed hybrid approach improves the sentiment classification accuracy.

G. Opinion Mining of Real Time Twitter Tweets

A. Shrivatava, S. Mayor and B. Pant [8], proposed "Opinion Mining of Real Twitter Tweets," In this proposed system, a tweet puller is developed which automatically fetch the public opinion on a topic and using SVM the opinions are classified into positive, negative and neutral. First, tweets are collected using twitter API then creating domain specific dictionary. Extracting all the tweets from Twitter when server is connected is done by tweet puller. Using classification tool to generate threshold frequency for each feature and generate a text file.

H. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis

L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu [9], proposed "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis". The system is a new entity-level sentiment analysis approach for Twitter, which is done using lexicon based method, the input preprocessed tweets are analyzed and categorize into sentence type detection, co reference resolution, using opinion rule aggregate opinions are formed which is input to train sentiment classifier that is Learning-based method and finally the extract opinionated tweets are classified. Co reference resolution gives the closest entity.

I. Twitter Sentiment Analysis The good, the bad and the neutral

Ayushi Dalmia [12] proposed "Twitter Sentiment Analysis The good, the bad and the neutral!". In this system, the lexicon based feature is further augmented by tweet specific features. The system includes English dictionary, acronym, and emoticon dictionaries. The preprocessing includes tokenization, removing non-English tweet, replace emoticons, remove URL, remove target mentions, remove punctuations from hash tags, handling sequences of repeated characters, removing numbers, removing nouns and prepositions, removing stop words, handle negative mentions and expand acronyms. After preprocessing & feature extraction, the tweets are feed into a classifier; it concludes that SVM gave the best performance. Hence, by building supervised system which merge lexicon based feature with tweet related features classify the tweets into 3-way classification-positive, negative and neutral.

IV. COMPARATIVE STUDY

This comparative study mainly based on above mentioned sentiment classification techniques. Table 1 shows studies done in different research papers on various classifiers and percentage accuracy given by those classifiers in year 2011 to 2015.

Table 1: Mining Techniques and their accuracy in different research papers

Studies	Mining Techniques Used	Performance (Accuracy)
A. Dalmia, M. Gupta, V. Varma [12]	Baseline Model	59.83%
A. Bakliwal et al. [10]	Baseline method on Stanford Dataset	87.2%
	Feature vector approach on Stanford Dataset	87.64%

A. Shrivatava, S. Mayor and B. Pant[8]	SVM	70.5%
A. Kumar, T. M. Sebastian[4]	Hybrid approach with Corpus based and dictionary based method	80%
J.I.Sheeba, Dr.K.Vivekanandan[6]	Fuzzy logic	85%
P. Chikersal, S. Poria, E. Cambria [11]	SVM	71.5%

From above study, Baseline method & Feature Vector Approach method on Stand ford Dataset found to be better mining technique for sentiment analysis. This system is improvised version of SVM (Support Vector Machine) that is semi supervised SVM classifier. It includes certain rules which can enhance the analysis mechanism by handling emoticons and punctuation, spell correction, stemming & stop word removal using unigram. In future, such system should be made to detect the problems like Sarcasm, irony, humor, mixed feeling, Named Entity Recognition (NER), Anaphora Resolution, conflicting signals, deliberate spelling mistakes, and Parsing. The Support Vector Machine has better accuracy when ruled based approach is used with it. The system should distinguish the text by analyzing whether the information is an opinion or just a fact. The domain sentiment analysis can be very efficient if amalgamated with other domains like fuzzy logic, speech recognition, and Artificial Intelligence.

V. CONCLUSION

This paper includes outline of current works that done on sentimental classification and analysis. From the survey we can conclude that supervised learning methods like Naïve Bayesian and Support Vector Machine are considered as standard learning method. Support Vector Machine provides excellent accuracy as compared to many other classifiers. In terms of accuracy we concluded that with small feature set Naive Bayes performs well, if large feature set is taken then

SVM will be the best choice. Lexical based approaches are ideally aggressive because it requires manual work on document. Maximum Entropy also performs better but it is suffered from over fitting. Fuzzy logic helps sentiment analysis provide efficient results as it is based on reasoning on the approximate values. Sentiment analysis when used with fuzzy logic helps to take decisions effectively but sometimes it may differ from the real time values. Many researches implemented opinion mining different techniques but still there is a need of automated analysis which addresses all the challenges of sentimental analysis simultaneously. A more innovative and effective techniques required to be invented which should overcome the current challenges like classification of indirect opinions, comparative sentences and sarcastic sentences.

REFERENCES

[1] Geetanjali S. Potdar, Prof R. N. Phursule, “A Survey Paper on Twitter Opinion Mining”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, Volume 4 Issue 1, January 2015, pp. 19-21.

[2] Md. Ansarul Haque1,Tamjid Rahman 2, “Sentiment Analysis By Using Fuzzy Logic”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4,No. 1,February 2014, pp. 33-48.

[3] VarshaSahayak, VijayaShete, ApashabiPathan, “Sentiment Analysis on Twitter Data”, International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 1, Volume 2, January 2015, pp. 178-183.

[4] Yakshi Sharma, VeenuMangat, And MandeepKaur, “Sentiment Analysis & Opinion Mining”, Proceedings Of 21 St Irf International Conference, 8 Th March 2015, Pune, India, Isbn: 978-93-82702-75-7, Pp. 35-38.

[5] Akshi Kumar and Teeja Mary Sebastian, “Sentiment International Journal of Computer Applications (0975 – 8887) The National Conference on Role of Engineers in National Building 24 Analysis on Twitter”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012, pp. 372-378.

[6] J.I.Sheeba and Dr.K.Vivekanandan, “A Fuzzy Logic Based On Sentiment Classification”, International Journal Of Data Mining & Knowledge Management Process (Ijdkp) Vol.4, No.4, July 2014, pp. 27-44.

[7] Namita Mittal, BasantAgarwal, SaurabhAgarwal, ShubhamAgarwal, Pramod Gupta, “A Hybrid Apprach for Twitter Sentiment Analysis,” Proceedings of ICON-2013: 10 thInternational Conference on Natural Language Processing, Noida, India, 2013, pp: 116-120.

[8] A. Shrivatava, S. Mayor and B. Pant, “Opinion Mining of Real Twitter Tweets,” International Journal of Computer Applications, Volume 100- No. 19, August 2014.

[9] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, “Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis,”Technical report, HP lLaboratories, 2011.

[10] AkshatBakliwal, PiyushArora, SenthilMadhappan, Nikhil Kapre, Mukesh Singh, VasudevaVarma, “Mining Sentiments from Tweets”, 3rd Workshop on Sentimentand Subjectivity Analysis (WASSA), Report No: IIIT/TR/2012/-1,July 2012.

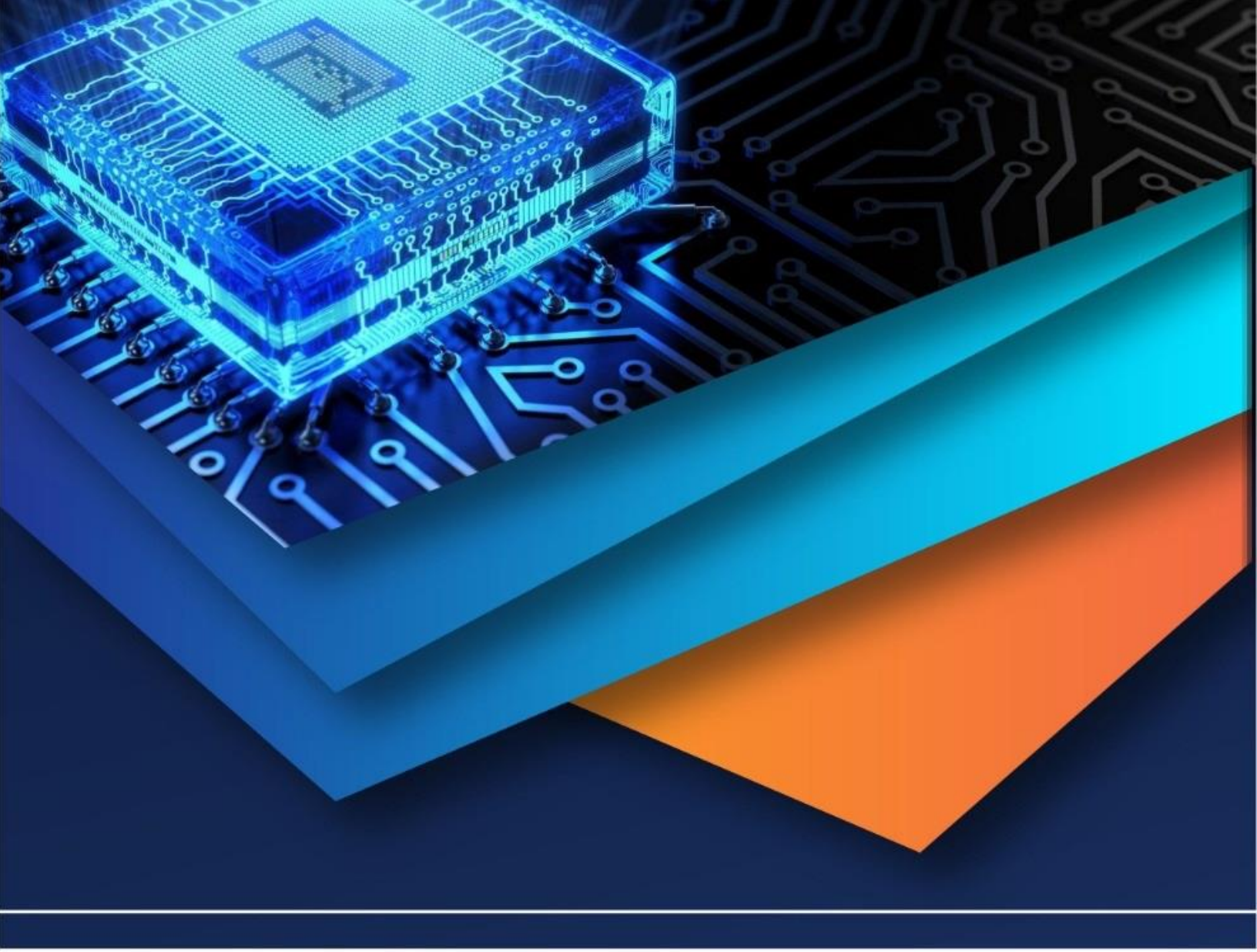
[11] PrernaChikersal, SoujanyaPoria, and Erik Cambria “SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning”, Proceedings of the 9th International on Workshop Semantic Evaluation (SemEval 2015), Denver, Colorado, June 4-5, 2015, pp. 647-651.

[12] AyushiDalmia, Manish Gupta_, VasudevaVarma, “IIIT-H at SemEval 2015: Twitter Sentiment Analysis The good, the bad and the neutral!”, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, June 4-5, 2015, pp. 520-526.

[13] EfthymiosKouloumpis, TheresaWilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!” Proceedings of the International AAAI Conference on Weblogs and Social Media, 2011, pp. 538-541.

[14] Maqbool Al-Maimani, NaomieSalim, Ahmed M. Al- Naamany, "Semantic and Fuzzy Aspects of Opinion Mining”, Journal of Theoretical and Applied Information Technology Vol. 63 No.2, 20th May 2014, pp. 330-342.

[15] Alexander Pak, Patrick Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, pp. 1320-1326



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY