International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

An Opinion Mining and Predicting Outcome for IPL Using Machine Learning Techniques

Miss. Amruta A. Gujar and N.G.Pardeshi

Department of Computer Engineering SRES SCOE,Kopargaon

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Nowadays, Social media used in many cases. It is also used to express the opinions or emotions on particular topic. There are various social media exist like facebook, twitter, linkedin etc. Now, Peoples having their account on more than one social media account which they can use for the purpose of sharing their thoughts, emotions etc. Twitter is mostly used social media account where people comment on tweets. In IPL system when someone make tweet related to the IPL. There are biggest fans of the IPL who comment on the tweet related to the IPL like share their emotions related to particular match or player. Tweet related to the IPL get extracted using the Phrases and hashtags . Machine learning technique random forest used to extract that tweets and perform some operation on that extracted tweet and classify it into various categories. Tweets get classified in the classes like Positive, Strongly positive, negative, strongly negative, neutral etc. To get the tweet from the twitter it need to access the twitter API using the credentials of twiiter account. For random forest method NLP library and the manual dataset used to verify the results. NLP is nothing but the natural language processing library where it contains set of words which is useful to match the extracted result.

Keywords: *Opinion mining, Machine learning, random forest, NLP, Twitter.*

Introduction

The Opinion mining is nothing but the deals with identifying and understanding opinions and sentiments expressed in a particular text. The masses give their opinion regarding various subjects on social media platforms using tweets, status updates and blogs. People can freely express their opinions and feelings on almost everything in a large crowd. Twitter is a standout amongst the most widely recognized online web based social networking and miniaturized scale blogging administrations which are an extremely well known strategy for communicating feelings and connecting with other individuals in the online world.
Human life is filled with emotions and opinions. An access to large amount of data through internet and its transformation into a social media is no longer an issue, as there are terabytes of new information produced on the web every day that are available to any individual. It now change the way to share the various information. The use of social media is increased day by day Increased growth of social media users on internet has also increased their participation in various discussions, events and activities simultaneously. In case of a things, reviews of various users will help to take many important decisions about the services of that particular thing. But manually reading such a large amount reviews is a very difficult task for persons. So there is a need of a system which will help to automatically extract the positive, negative and neutral features of the product and make the decision making process easier. There are many sites and companies which perform these activities[6].

Twitter messages give genuine crude information in the organization of short messages that express sentiments, thoughts and occasions caught at the time. Tweets and messages are short: a sentence or a feature as opposed to an archive. While there is no restriction to the scope of data passed on by tweets and messages, regularly these short messages are utilized to impart insights and opinions that individuals have about what is going on in their general surroundings and in peoples life. The dialect utilized is exceptionally casual, with inventive spelling and accentuation, incorrect spellings, slang, new words, URLs, and shortenings, for example, RT for "re-tweet" and # hash labels, which are a kind of labeling for Twitter messages. Another part of web-based social networking information, for example, Twitter messages is that it incorporates rich organized data about the people required in the correspondence. For instance: Twitter keeps up data of who takes after whom and retweets and labels within tweets give talk information. When taken in accumulation tweets can give an impression of open assessment towards

occasions, proposed framework give a positive or negative estimation on Twitter posts utilizing an outstanding machine learning strategies for content order, called

Bayesian Logistic Regression [BLR] and Naive Bays Classification for giving positive or negative conclusion on tweets.

We look at one such popular social media called Twitter and build various models for classifying "tweets" into positive, strongly positive, strongly negative, negative and neutral sentiment. Twitter is one of the most concise and precise microblogging sites. Individuals always tweet over latest happenings with a "Hashtag" which is frequently the fundamental subject for examination. The data is collected using the streaming API of twitter. The benefit of this data on manually used data sets or NLP libarary is that the tweets are collected in a streaming fashion and therefore represent a true sample of original tweets in terms of languages use and data. Sentiment Analysis on a sentence-level can be done using lexical approach. Vader, an open source tool and a part of NLTK python library, carries out the part of the sentiment analysis on a sentence-level. A method named Sentiment Intensity Analyser under Vader is useful to calculate the polarity of the text and classify them accordingly among positive, negative or neutral.

Using social media like twitter, models are built for classifying "tweets" into positive , Strongly positive, Strongly negative, negative, and neutral classes .The models are build for two classification tasks : a 5-way classification of already separated phrases and hashtags in a tweets into positive, Strongly positive, strongly negative, negative , and neutral classes and another 5 way classifications of entire message into positive, strongly positive, strongly negative , negative and neutral classes.

Tweets are extracted using the program written in python programming language. After the extraction of tweets, the natural language processing tasks are carried out. Natural language processing is the ability to make the computer to understand the everyday language (English language). In natural language processing, different functions are performed such as part of speech tagging, term frequency and independent document frequency, feature computation. Then, machine learning algorithms are used. Basically there are two types of machine learning techniques. They are supervised and unsupervised machine learning technique. Classification is performed in supervised learning technique where the machine learns from training set of data. But in unsupervised machine learning technique, the machine learns from unknown labels of data. The proposed project is based on supervised machine learning techniques such as naive Bayes classifier, support vector classifier, logistic regression. The accuracy of these machines learning classifier is computed and finally overall accuracy is calculated for our opinion mining[4].

In the next section, we make a review of algorithms that handle large dimensional imbalanced data. In section 3 we introduce system architecture which shows the flow of system and section 4 describes system analysis. Paper is concluded in section 5.

Literature Survey

In recent years, a branch of machine learning which is become more and more popular and also it has been applied for opinion mining; previous studies have shown that machine learning has potential to create better model than traditional methods.

In Arti, Kamanksha Prasad Dubey, Sanjay Agrawal[1] proposed system, to Analyze the opinions for sporting game such as IPL 2016 Twitters API give access to authorized users to extract the details insights into tweets post. The practical result displays the positive think and negative think of individuals respectively. This type of an opinion analysis might offer valuable feedback to the business and facilitate them to identify a negative tweets flip in viewer's understanding. Deciding negative trends too soon will permit them to form educated choices on a way to target specific aspects of their services and products so as to extend its client satisfaction. They illustrates the views of 'Random Forest' that has primary result on comprehensive accuracy of the interpretation.

In D. Krishna Madhuri[2] presented a methodology for sentiment classication has proposed. The case study considered is Indian Railways. It explores supervised learning methods like C4.5, Naive Bayes, SVM and Random Forest in sentiment classication of tweets of Indian Railways. Since tweets carry valuable social feedback and opinion on Indian Rail- ways, this study provides useful insights on sentiment classication. It considers positive, negative and neutral sentiments.

Mariam Adedoyin-Olowe ,MohamedMedhat Gaber and Frederic Stahl[3] proposed system that analysing SM data especially opinions/sentiments expressed by SM users with data mining techniques has proved effective and useful considering the research carried out so far in the field. This is because of the limitation of the data mining possess in handling noisy, large and manual data. Different writers have introduces and tried several algorithms that can be used to extract the opinions of online users of the SM. More number of works reviewed and it observer that majorly utilized Support Vector Machine (SVM), Naive Bayes and Maximum Entropy.

Ali Hasan , Ahmad Karim and Shahaboddin Shamshirband [4] define the issued of more than one sentiment analyzers with machine-learning algorithms to determine the approach with the possible accuracy rate for learning about election opinions. In a lexicon-related sentiment analysis ,semantic view of words, different phrases or sentences calculated in a document. Polarity in the lexicon-related method is calculated on the basis of the dictionary, that consists

of a semantic score of each and every word. However, the approach of machine learning is basically destined to classify the text by applying algorithms such as Naıve Bayes and SVM on the les.

Er. Hari K.C. [5] proposed a task in simply predicting the popularity of two biggest brand smartphones based on the platform they used. The varieties of Machine learning approaches are used to analyze the tweets from twitter related to smartphones.

Grigori Sidorov , Sabino Miranda-Jimenez , Francisco ViverosJimenez [6] presented an analysis of various parameter settings for selected classiers: Supported Vector Machines, Naıve Bayes and Decision Trees. We used n-grams of normalized words (additionally ltered using their POS-tags) as features and observed the results of various combinations of positive, negative, neutral, and informative sets of classes. We made our experiments in Spanish language for the topic related to cell phones, and also partially used data from tweets related to the recent Mexican presidential elections (for checking the balanced vs. unbalanced corpus).

Vidushi , Gurjot Singh Sodhi [7] proposed a methodology for the classication of sentiments was developed in this paper for reviews on purchased items in Indian market.12 product reviews were used in this paper. The streamed reviews was ltered for relevant content and stored in a database. The several steps of pre-processing were applied on it and the reviews were removed from special characters, stop word, tokenized, etc.

Kalaivani A ,Thenmozhi D [8] proposed a literature survey on the different DL techniques associated with SA. The SA importance is also delineated. In addition, the disparate types of classication process and their limitations are discussed briey. This literature work enlightens the various prevailing methods of SA proposed by diverse researchers, which assist the forthcoming researchers in this specic area.

Ghazaleh Beigi , Xia Hu , RossMaciejewski and Huan Liu [9] defined system that covered state-of- the-art sentiment analysis approaches and show their involvement and then discussed the application of social medias and sentiment analysis in disaster relief and situational awareness, while they also detailed applications of visual analytics with an emphasis on sentiment analysis. In this section they discuss some of the challenges facing the studies in sentiment analysis and its application in disaster relief, as well as visual analytics.

Soujanya Poria , Erik Cambria, Alexander Gelbukh [10] introduced the rst deep learning-based approach to aspect extraction. As expected, this approach gave a signicant improvement in performance over state-of-the-art approaches. They proposed a specic deep CNN architecture that comprises seven layers: the input layer, consisting of word embedding features for each word in the sentence; two convolution layers, each followed by a max-pooling layer; a fully connected layer; and, nally, the output layer, which contained one neuron per each word. They also developed a set of heuristic linguistic patterns and integrated them with the deep learning classier.

Ritu Mewari, Ajit Singh, Akash Srivastava [11] proposed system that exist a lot of benefits of opinion mining at customer and business level. Opinions study about particular product provides us a accurate picture of future. So company can modify their product according to customers need and customer can aware about that particular product before going to purchase it. A amount of data is daily posted on social media like facebook ,twitter e.t.c. User says their opinion in the form of comments, reviews, tweets and feedback daily .An opinion mining process gives us the way to extract pearl knowledge from it.

David Osimo and Francesco Mureddu [12] focusing on improving the accuracy of algorithm for opinion detection. Semantic analysis through lexicon/corpus of words with known sentiment for sentiment classication.It also identify of highly rated experts.

Proposed Methodology

Random forest, which were formally proposed in 2001 by Leo Breiman and Adèle Cutler, are part of the automatic learning techniques. This algorithm combines the concepts of random subspaces and "bagging". The decision tree forest algorithm trains on multiple decision trees driven on slightly different subsets of data.

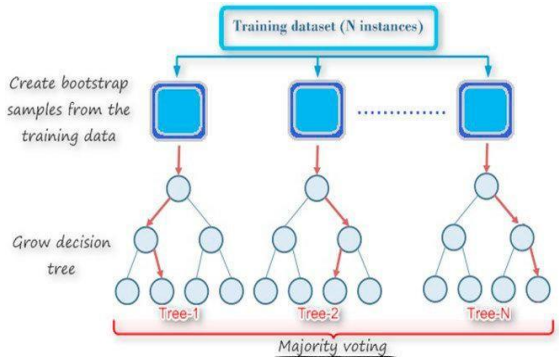


Fig 1. Pictorial representation of random forest

The random forest is part of the family set methods that take the decision tree as an individual predictor, they are based on the methods of Bagging, Randomizing Outputs and Random Subspace excusing boosting. The random forest algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy. It is an ensemble learning method for classification and regression that constructs a number of decision trees at training time and delivers the class that is the mode of the classes output by individual trees.

Random Forest Algorithm:

- For b = 1 to B Make
- Draw a sample point Z* of size N from the given training data.

• Draw a random forest tree Tb to the bootstrap data, by recursively calling the following steps for each terminal node of the tree, until the minimum node of the size nmin is reached.
• Select m variables from the list the p variables.
• Pick the best variable and split the data among the m.
• Split the nodes into 2 daughter nodes.
• Output after the ensemble of trees . $T_{B_1}^{B}$

To make a predict data at a new point x:

Regression: $\hat{f}_f^B(x) \Leftarrow \frac{1}{R}\Sigma_{b=1}^{B} T_b(x)$

Classification: Let take $\hat{c}_b(x)$ be the class prediction of the both random forest tree.

Then $\hat{C}_{T_f}^{B} — majority\ vote\{\hat{c}_b(x)\}B$

In random forest classification method, many classifiers are generated from smaller subsets of the input data and later their individual results are aggregated based on a voting mechanism to generate the desired output of the input data set. This ensemble learning strategy has recently become very popular. Before RF, Boosting and Bagging were the only two ensemble learning methods used. RF has been extensively applied in various areas including modern drug discovery, network intrusion detection, land cover analysis, credit rating analysis, remote sensing and gene microarrays data analysis etc.

There are two ways to evaluate the error rate. One is to split the dataset into training data and check the part. We can check the training part to construct the forest, and then use the check part to measure the error rate. Another way is to use the Out of Bag (OOB) error estimate. Random forests algorithm measures the OOB errors in the the training phase, we do not need to split the training data.

The architecture of the system is shown in figure2.Twitter API need to access the tweets given by the user. Twitter api access by using the credentials which is confedential. After tweets collection preprocessing done on that tweets using hashtag and phrases. Feture get extracted from that tweets.After extraction data machine learning algorithm such as random forest applied on that extracted data. Then data get classified into different classes.

Use of random forest algorithm:

A) Random forest algorithm can be used for purpose like the classifiaction and the regression tasks. B) It provides higher accuracy.
C) Random forest classifier will handle the missing values and it maintain the accuracy of a large proportion of data.
D) If there are more trees, it won't allow overfitting trees in the model.
E) It has the power to handle a large data set with higher dimensionality..

A .Twitter Streaming API:

API stands for Application Programming Interfaces (APIs) and that allows you to access resources which available on the server. Lets learn how to use twitters API. ... You will then need to go to apps.twitter.com and create app so we can reference the coresponding keys Twitter generates for this app .To access twitter api some steps need to be performed.Firstly,open twitter account goes to the setting. Click on developer option. Click on edit where you get the access token, access secret ,consumer secret ,consumer token key.This keys are necessary to access the API of that particular account.

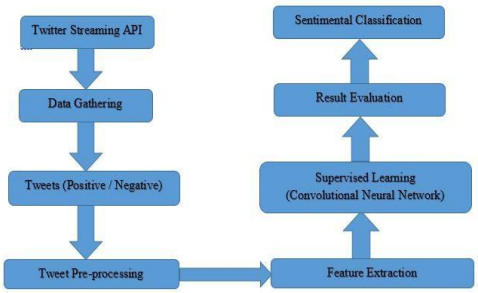


Fig 2. Proposed system architecture *B .Data gathering:*

Data gathering include the collection of all data related to the IPL. This data gathered from the phrases and the hashtags given by the user. Data gather in one set.

C Tweet Preprocessing:
Tweet preprocessing includes the tokenization ,stemming and stop words removal. Preprocessing helps to remove all the garbage data from the tweets. It separate each word and removes that word which doesn't mean for the tweets. It also removes the a, an, the fro the senetencs. It removes the blank spaces, URL, tabs from sentence which takes lots memory.
D Feature Extraction:
Feature extraction contain the extraction of the words from the tweet like phrases and hashtags. It extract the sentiment related words from the tweets. In the case of IPL, the words get extracted from the hashtag such as the #IPL, #happy etc. After feature extraction the random forest algorithm applied on the extracted word and final result get calculated according to the various sentiment classes.
E. Dataset Description(NLP):
NLP is nothing but the natural language processing library which is used for the interactions between computers and human (natural) languages, in particular how to process computers to process and analyze large amounts of natural language data. Natural language processing or NLP is a field of machine learning that focuses on enabling machines to understand and interpret human languages just like the programming languages.

System Analysis

A. Mathematical model:

Set theory is used to present the mathematical model of the system, which includes different components like input, rule, process, and output.

Based on the set theory the relevant mathematical model is

described as follows,

S = { I, P, R, O }

Where,

- **S** is the System for An opinion mining and predicting outcome for indian premier league using machine learning techinque.
- I is set of Initial Input to the system.

I = { I1, I2 } I1 = Training Tuples.

I2 = Testing Tuples.

- P is set of procedure or function or processes or methods.

P = { P1, P2, P3, P4, P5 }

P1 = Process for reading dataset.

P2 = Process for Data Preprocessing.

P3 = Process of Feature Selection.

P4 = Process of Training.

P5 = Process of Testing.

- R is a set of rules.
- O is a set of outputs.

O = *{ O*1, *O*2 *}*

*O*1 = Correctly Classifified Class. *O*2 = Incorrect Classifification.

- **Venn Diagram :**

Venn diagram represents the interaction between different processes along with input and output.Fig.3 shows the mapping of input,process and output.where *I*1, *I*2 are inputs, *P*1, *P*2 , *P*3, *P*4, *P*5 are process and *O*1, *O*2 are Outputs.

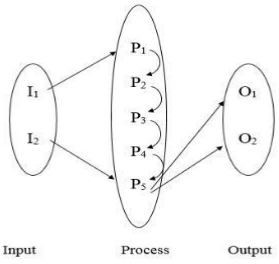


Fig 3. Venn diagram

B. Experimental evaluation: 1 Evaluation: The performance of opinion classification can be evaluated by using four indexes calculated as the following equations:

$$\text{Accuracy} = \frac{TP+TN}{FP+FN+TN+TP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

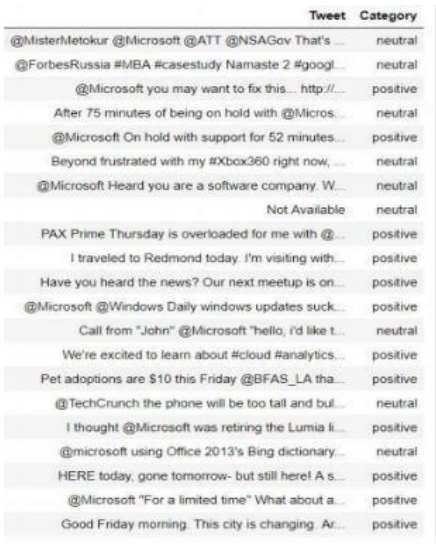
F1=(2×Precision×Recall)/(Precision+Recall) Where, λ True Positive (TP) means correctly rejected i.e., the number of anomaly records that are identified as anomaly. λ False Positive (FP) means incorrectly rejected i.e., number of normal records that are identified as anomaly. λ True Negative (TN) means correctly admitted i.e., the number of normal records that are identified as normal. λ False Negative (FN) means incorrectly admitted i.e., number of anomaly records that are identified as normal.

Table 1. Confusion Matrix

	Predicted positives	Predicted Negatives
Actual Positive	TP	FN
Actual Negative	FP	TN

2. Results

We used the twitter dataset publicly made available by Stanford university. Analyses was done on this labeled datasets using various feature extraction technique. We used the framework where the preprocessor is applied to the raw sentences which make it more appropriate to understand. Further, the machine learning technique trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content. Here, NLP dataset is used for the opinion mining NLP is nothing but the natural language processing.It is contain the collection of words. The Random Forest classifier model is implemented on the dataset and the accuracy of prediction of sentiment analysis on the dataset is found to be 56.7%. The Metric values obtained from the experiment are presented in Figure 4.



Tweet	Category
@MisterMetokur @Microsoft @ATT @NSAGov That's ...	neutral
@ForbesRussia #MBA #casestudy Namaste 2 #googl...	neutral
@Microsoft you may want to fix this... http://...	positive
After 75 minutes of being on hold with @Micros...	neutral
@Microsoft On hold with support for 52 minutes...	positive
Beyond frustrated with my #Xbox360 right now, ...	neutral
@Microsoft Heard you are a software company. W...	neutral
Not Available	neutral
PAX Prime Thursday is overloaded for me with @...	positive
I traveled to Redmond today. I'm visiting with...	positive
Have you heard the news? Our next meetup is on...	positive
@Microsoft @Windows Daily windows updates suck...	positive
Call from "John" @Microsoft "hello, i'd like t...	neutral
We're excited to learn about #cloud #analytics...	positive
Pet adoptions are $10 this Friday @BFAS_LA tha...	positive
@TechCrunch the phone will be too tall and bul...	neutral
I thought @Microsoft was retiring the Lumia li...	positive
@microsoft using Office 2013's Bing dictionary...	neutral
HERE today, gone tomorrow- but still here! A s...	positive
@Microsoft "For a limited time" What about a...	positive
Good Friday morning. This city is changing. Ar...	positive

Fig 4. Random forest classifier prediction

The opinion mining on the particular team can be shown in the graphical view as below:

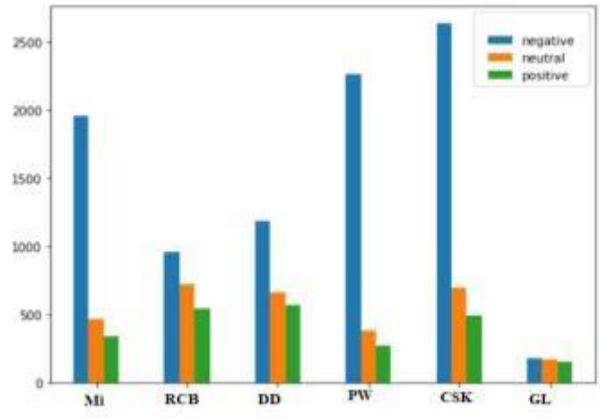


Fig 5. Opinion mining on IPL

We have found some significant improvement in accuracy. Table II contains proposed work accuracy compared with SVM algorithm. Fig 6 contain accuracy comparison with a different algorithm using a bar chart. As an analysis, Random forest method gives an overall 95.6% accuracy.

Table 2. Compare proposed accuracy and existing accuracy

Algorithm	Accuracy
Support vector machine (SVM)	78.4%
Random Forest	95.6%

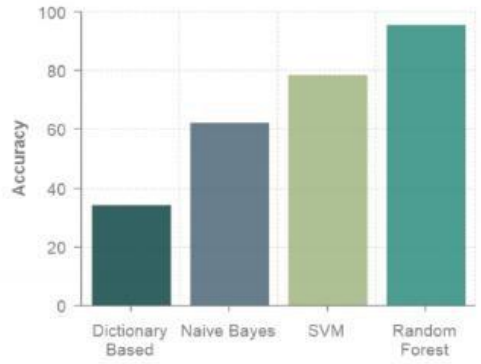


Figure 6. Different classifier accuracy

Conclusion

The IPL opinion mining using random forest helps to extract data from the twitter account. Data extracted from the tweet using the hashtags and phrases.It also shows the performance of perticular player according to the strike rate. It helps to show all history of players.The NLP dataset and mannual dataset used to match the results and classifies the data into the different classes such as the strongly positive, positive, neutral, strongly negative, negative etc. Classification into appropriate classes done by using the random forest algorithm. which filters all data. There are also others mehods in machine learning for opinion mining like SVM, naive bayes, CNN etc. Random forest gives the more appropriate result. ACKNOWLEDGEMENT I would like to thank my esteemed guide, Prof. N.G.Pardeshi, whose guidelines inspired to work. Dr. D.B. KShirsagar H.O.D who has bought enthusiasm to explore the topic.

References

[1] Arti, Kamanksha Prasad Dubey and Sanay Agrawal-*Opinion mining for indian premier league using machine learning techniques* || 978-1-7281-12534/19/31.00 c 2019 IEEE.

[2] D. Krishna Madhuri-*A machine learning based framework for sentiment classification: indian railways case study* || International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075, Volume-8 Issue-4, February 2019.

[3] Mariam Adedoyin-Olowe , Mohamed Medhat Gaber and Frederic Stahl-*A survey of data mining techniques for social media analysis* || School of Computing Science and Digital Media, Robert Gordon University Aberdeen, AB10 7QB, UK.

[4] Ali Hasan , SanaMoin , Ahmad Karim and Shahaboddin Shamshirband,m*achine learning-based sentiment analysis for twitter accounts* || Mathematical and computational applications .,2018, 23, 11.

[5] Er. Hari K.C, *"Online social network analysis using machine learning techniques* || International Journal of Advances in Engineering Scientifific Research, Vol.4, Issue 4, Jun-2017.

[6] Grigori Sidorov , Sabino Miranda-Jimenez , Francisco Viveros-Jim ´ enez , Alexander Gelbukh , Noe Castro-S ´ anchez , Francisco Vel ´ asquez , Ismael Dıaz-Rangel , Sergio SuarezGuerra , Alejandro Trevino and Juan Gordon *Empirical study of machine learning based approach for opinion mining in tweets* || Center for Computing Research, Instituto Politecnico Nacional. ´

[7] Vidushi and Gurjot Singh Sodhi-*Sentiment mining of online reviews using machine learning algorithms* || Department of Computer Science Shaheed Udham Singh College Of Engineering Technology,2017 IJEDR — Volume 5, Issue 2 — ISSN: 2321-9939.

[8] Kalaivani A and Thenmozhi D-*Sentimental analysis using deep learning techniques* || Blue eye intelligence engineering and science publication.,December 22, 2018.

[9] Ghazaleh Beigi , Xia Hu , Ross Maciejewski and Huan Liu - *An overview of sentiment analysis in social media and its applications in disaster relief* || Computer Science and Engineering, Arizona State University,Department of Computer Science and Engineering, Texas AM University.

[10]Soujanya Poria , Erik Cambria and Alexander Gelbukh-*Aspect extraction for opinion mining with a deep convolutional neural network* || Temasek Laboratories, NanyangTechnological University, Singapore,Knowledge-Based Systems 108 (2016).

[11]Ritu Mewari, Ajit Singh, Akash Srivastava, *"Opinion Mining Techniques on Social Media Data"*, International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 6, May 2015.

[12]David Osimo and Francesco Mureddu-*Research challenge on opinion mining and sentiment analysis* || Anderson, C. (2008). Wired Magazine, 16(7), 16–07.