An Opinion Mining And Predicting Outcome For Indian Premier League Using Machine Learning Technique

Miss.Amruta Gujar[1], Prof.N.G.Pardeshi[2]
[1]Dept of Computer Engg,
[2]Associate Professor, Dept of Computer Engg,
[1, 2] Sanjivani college of engg,Kopargaon

Abstract- *With the increasingly use of the internet nowadays, Social media becomes more important part of the human being life.There are multiple social media existed such as the twitter, facebook, linkedin etc.People analyze and share their personal opinion and thoughts about the particular event or things on social media such as opinion related to IPL matches. Social media like twitter used to collect the opinions of peoples regarding to the IPL matches. It collect the opinion from the tweets and distribute into the classes. It categories in the classes by the hashtag and phrases which is used by the peoples while tweets. It classifies the opinions using the machine learning algorithm such as the random forest. Random forest algorithm have the more accuracy and gives the most accurate result compares to the SVM, naive bayes algorithms.It uses the dataset to compare and classifies the data into various classes. The thoughts classified in the different classes such as the positive, strongly positive, negative, strongly negative and neutral. It also shows the winning prediction.*

Keywords- Social network, Twitter API, Random Forest, Machine Learning, NLP

I. INTRODUCTION

Social media plays a vital role to share content and public opinion among the communities. It is a tool that can instantly and actively embrace the publics dialogue. Social media has turned into a dynamic stage for the web users to share their views about social activities, sports events and political occasions and so forth. Subsequently, social media contains amonstrous amount of information that can be utilized for supposition mining and wistful investigation. On certain social and political issues, governments and NGO's seek rapid input of the masses; and in this regard, social media covers a convenient solution. Individuals can candidly participate in dialogs about an event e.g,, political situation in a country, general elections, terrorist attacks, sports events, release of a movie, IPL matches and so on. Advancements in data analysis have expanded social networks. Social media is a stage where individuals talk about and share data unreservedly. The verity of occasions are celebrated on online networking, for example, Twitter, Facebook, Google+ and so forth. Social media events are classfied as arranged events and spontaneous events.

Most of the people uses the social media, even peoples stay updated by uses of the social media. From sharing the e motion ,thoughts to spread any information to the people social media is mostly uesd. Twitter is most active social media which helps to share opinion of various people on different events. People express while they gives opinion on particular topic. In twitter ,while tweets most of the person uses the hashtags and phrases at the time of the commenting. This phrase and hashtag plays a very important role to extract the tweets on any particular topic like IPL. To extract that tweet firstly need to log in into the account.

Login helps to access the credentials of that account. Firstly, it need to access twitter API. After accessing the twitter API user need to log in into that account using username and password. After log in into the twitter account to access that account tweets it need to take the credentials of that account.For that purpose we need to create the apps of account.after creation of app we need to go to the developer option.In developer option click on the app.In this we get the credentials of that account which remains confedential.

Opinions, speculation, emotions and evaluation typically reveals the views of different from classs; they include opinionative narrow data express during a language which is compiled of subjective statements . "Sentiment analysis task is to identify people's belief, assessments, thought, appraisals, sentiment, and feelings towards instances like services, outcomes, events,activities ,features and their framework.". paper provides information about public's perceptions towards an event IPL 2016. [1].

Machine learning algorithm used to classified the data into the various classes.Random forest algorithm used to evaluate the data into the classes.

II. LITERATURE REVIEW

In order to predict the opinions for sports game such as IPL 2016 Twitters [1] API give access to authorized users by using the credentials to extract the details into tweets post. The result shows the positive way of thinking and negative way of thinking of that user respectively. This type of an opinion analysis helps to provide valuable feedback to the organization and helps to them to recognize a negative tweets flip in users point of view. Deciding negative trends too easily will allows users to make educated choices on a way to target particular feature of their service and product so as to improve its client satisfaction. In this paper defines the views of ‗Random Forest' that has firstly result on their accuracy of the prediction or analysis. This proposed methodology has an accuracy of 81.69% for classification of tweets. So the comparison between the totally different techniques and proposed method shows that the proposed techniques is most preferable in the essential analysis of the measurable attributes of exactly classifying out of all examples (tweets), specificity, F-score and the Area Under Curve.

A methodology for sentiment classification [2] 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 classification of tweets of Indian Railways. Since tweets carry valuable social feedback and opinion on Indian Railways, this study provides useful insights on sentiment classification. It considers positive, negative and neutral sentiments. The proposed framework has two phases such as training and testing. In the training phase, the classifiers are built to have different knowledge models that can be used in sentiment classification. Analysing SM [3] data especially opinions/sentiments defines by the SM users with the data mining techniques has proved more effective and useful considering the research done in it very deeply in that particular field. This is so because of the capacity data mining helps in managing noisy, large and dynamic data. Different authors have introduce the several algorithms and done study in it ,that can be used to fetch the opinions of the online users of the SM. Many number of works reviewed majorly utilized Support Vector Machine (SVM), Naive Bayes and Maximum Entropy.

The adoption of different sentiment analyzers [4] with the machine-learning algorithms to identifies the approach with the highest accuracy rate for learning and to get the most appropriate result about the election related sentiments. In a lexicon-based sentiment analysis, semantic orientation is of the words, phrases or various sentences measured in a document. Polarity in the lexicon-based method is calculated on the basis of the words in the dictionary, that consists and shows the semantic score of a particular word. However, the way of machine learning is basically destined to categorized the text by applying the different algorithms such as Naïve Bayes and support vector machine on the files. Most of the worked is done in this field considering the various aspects of the sentiments and the lexicon using the approach of machine learning.

The set of information [5] includes in Twitter helps to extract the data very appropriately, it makes an attractive source of data for opinion mining and sentiment analysis. After calculating the overall performance of the machine learning algorithm it conclude that performance is good when classifying sentiments in the form of tweets, both in English and in Spanish language. We comes to know that the precision can be further improved using more used features. In this research, we calculated an analysis of the various parameter settings for selected classifiers: Supported Vector Machines, Naïve Bayes and Decision Trees. We used n-grams of normalized words as features and observed the results of various collection of positive, negative, neutral, and informative sets of various classes. We made our experiments in Spanish language for the topic related to mobile phones , and also used data from tweets related to the recent elections .

From the overall observation of the all results, we found that the best configuration of systems was: (1) using the unigrams model as the features,(2) using less predictably number of classes: positive and the negative, (3) using at least 3,000 tweets as training set (4) managing the corpus as related the proportional representation of all the classes gives somehow bad results, and (5) Supported Vector Machines was the classifier with the best precision and accuracy.

A methodology for the classification [6]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 modified after applying the filter to the data for related data and stored in a database. The various steps of pre-processing were applied on it and the final reviews were removed out from special characters, stop word, tokenized, etc. Stemming was completed to all words in order to extract features and the root words from it. TF-IDF score based approach was developed and the score was calculated for each separate reviews. Feature Selection was processed on it using Chi Square method and information gain. The extracted feature from the sentences forms a term document matrix which is evaluated in the classification algorithm.

III. SYSTEM OVERVIEW

The opinion mining system for IPL matches using the random forest algorithm is shown in the below figure. In the system firstly user need to take the twitter API to access the twitter. User need to enter the credentials of user account which remain confedential. After that data gathering helps to collect the data from the twitter. After collecting data which contain mixed data that is positive, negative and neutral etc. It perform preprocessing on collected data like removing stop words, stemming. After that feature extraction performs extraction of the features by hashtag and phrases. Random forest algorithm used to perform sentiment analysis. Finally,result of opinion mining get evaluated..

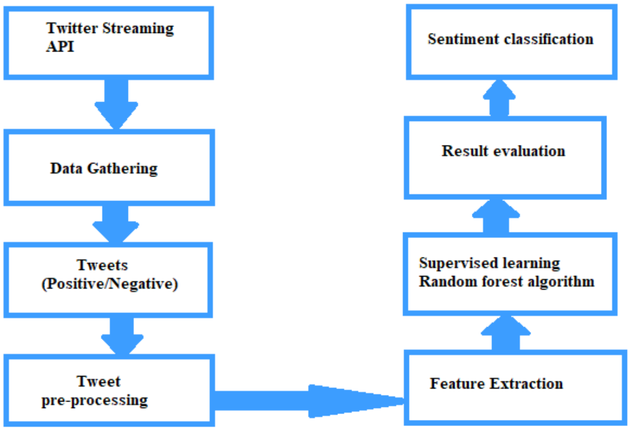


Fig.1Proposed System Architecture

We will use the typical machine learning methodology. We will first add the dataset and then we will then do analysis on this to look that is we can find any trends in the dataset. Next, we will apply the text preprocessing to convert textual form data to numeric form data that can be helpful to machine learning algorithm. So, here machine learning algorithm used to train and test the sentiment analysis of the IPL matches from twitter.

A . Dataset:

The dataset here used to evaluate the system is taken from the kaggle.com. This dataset includes data related to the IPL matches. Dataset include the classes that is Strongly positive, positive, neutral, strongly negative, negative etc.Dataset contains the collection of words which categorizes into different classes. Data stored in various classes helps to identify the type of tweet. It helps to recognize thoghts of the people about the IPL matches. It also helps to decide the winning prediction of match. Prediction decided according to the sentiment classes such as the if tweet is strongly positive then winning prediction is 80% to 100%,if its positive then winning prediction is 80% to 60%, Neutral class gives winning prediction up to 60% to 40%,Negative class shows predition up to 40% to 20% and strongly negative gives prediction up to 20% to 0%.

B.Preprocessing:

Preprocessing is very important procedure in opinion mining after extract the tweets. Preprocessing helps to remove the stop words from the sentance.It also corrects the spelling mistakes that present in the sentence. Preprocessing also helps to clean our dataset. It is very necessary to clean the tweet before going for training. Preprocessing useful to predict the actual class of the tweet. Preprocessing also unimportant data like the URL, a, an, the etc. It processes the data in the way to get the actual result of the IPL matches. Preprocessing includes the tokenization and stemming.

Tokenizing, tokenizing or parsing stage is the arrange the stage of the input string based on the each word arrange. In principle, this is nothing but the separate each and every word that related to a document. In general, each word is recognized or separated by another word by a space character, so the tokenizing process depends on the space character in the document to do word separations from each other word

.

Stemming is the stage to prepare the word initials into basic words. In stemming, conversion of morphological forms of a word to its stem is done considering each one is very related. The stem need not be an already exists word in the dictionary but all its content should appear to this form after the stemming has been completed. There are two points to be considered while using a stemmer :

- Morphological is the collection of a word are consider to have the same related meaning and hence should be merged in to the related stem.
- Words and characters that do not have the same mean should be kept isolated.
- These two rules are helpful enough as long as the final stems are useful for our opinion mining or language processing applications.

C. Random Forest Algorithm:

Random forest algorithm working shows in following diagram 2.Random forest algorithm is the supervised learning algorithm. It contain the collection of the bagging and boosting to perform the operation. It is collection of the classification and regression. Random forest algorithm construct set of classifier instead of single classifier and classify new data points by taking vote of their prediction. It is also called as ensemble learning approach. Ensemble learning means where you join different types of algorithm or the same algorithm multiple times to form a more powerful prediction model. It combines multiple algorithm of the same type which result in the forest tree that why it is called as the random forest algorithm. Random forest algorithm gives most accurate result compare to the other machine learning algorithm. It uses bagging and boosting in which bagging gives less accurate result than the boosting because it applies one model once. In boosting it continues the model untill they get the accurate result.

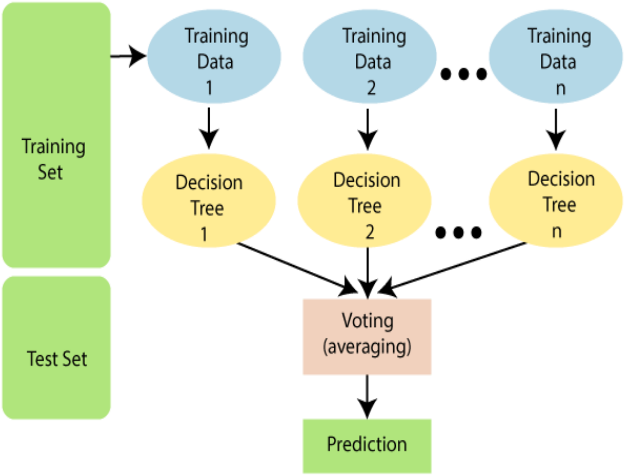


Fig. 2 Random Forest Algorithm

D. Training:

The training data is an initial set of data used to help a program understand how to apply technologies like neural networks to learn and produce sophisticated results. It may be complemented by subsequent sets of data called validation and testing sets. The generation of training models and test took place with the help of the WEKA tool. Through this, we used the implementations of algorithm Random Forest for classication, necessary for the creation of models.

E. Testing:

This is the final step where model performance is evaluated. In the csv datasets,the set is used for predicting the performance of model. Confusion matrix can be used to check the accuracy of model. Accuracy is most important performance indicator used to measure performance of random forest algorithm .

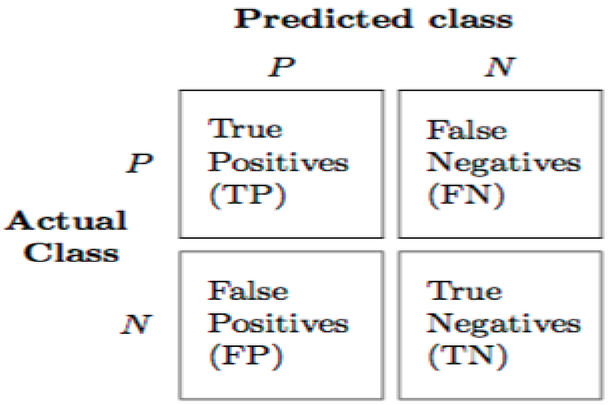
IV. SYSTEM ANALYSIS

A. Evaluation Metrics:

Accuracy is nothing but most important performance indicator of the system. The confusion matrix is used to find accuracy as shown in below table 1.It contain the TP,TN,FP,FN.

- TF means model correctly predicts negative class.
- FP means model incorrectly predicts positive class.
- FN means model incorrectly predicts negative class.
- TP means model correctly predicts positive class.

Accuracy is nothing but the number of records correctly classified into the total records.



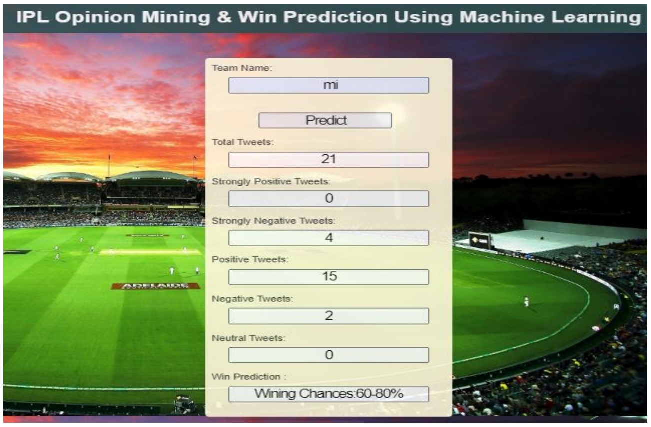
		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Fig.3 Confusion matrix

B. Results:

Result of the opinion mining system for IPL as shown below. In this we used random forest algorithm which is written in python. A personal laptop, which has a configuration of AMD E1-6010 APU with AMD Radeon R2 Graphics @ 1.35 GHz, 4 GB memory is used to perform this experiment and GPU acceleration is not used.

In the result when user enters the team name of IPL such as mi,rcb,kkr,dd etc.Then user need to click on the predict button after click on the predict button it shows the bar graph which is shown in figure 5.In the bar graph we shown the result for the mumbai indian team like mi. It shows the strongly positive,positive,strongly negative,negative,neutral count in bar graph. This count also display in the window in figure 4.In figure 4 it also shows the total number of tweet extract regarding to mi. It also shows the winning prediction of that particular team according to tweets classes in percentage. Like this it calculate the opinion for all IPL team and predict their winning prediction as shown in below figure 4.



IPL Opinion Mining & Win Prediction Using Machine Learning

Team Name: mi

Predict

Total Tweets: 21

Strongly Positive Tweets: 0

Strongly Negative Tweets: 4

Positive Tweets: 15

Negative Tweets: 2

Neutral Tweets: 0

Win Prediction : Wining Chances:60-80%

Fig. 4 Total tweet collected and winning prediction of MI

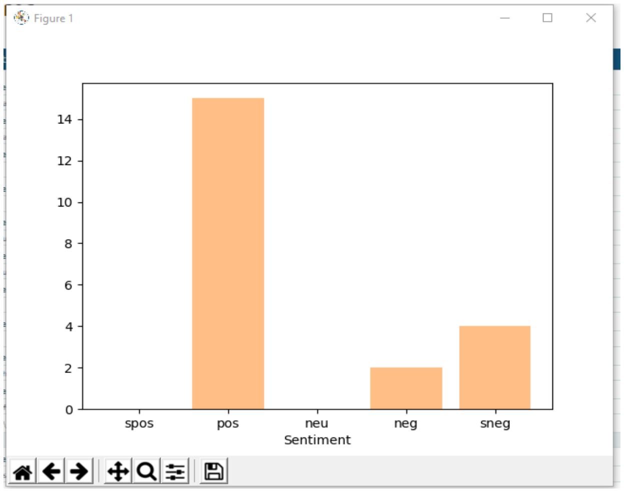


Fig. 5 Bar graph for mumbai indians

V. CONCLUSION

An opinion mining and predicting outcome for Indian Premier League using machine learning technique has ability to detect exact opinion of the user related to the IPL matches. It also shows the winning prediction of that particular team according to the opinion of users. Opinion mining is performed to obtained actual tweets meaning and distribute it into classes like Strongly positive,positive,Strongly negative,negative,neutral etc.It uses machine learning technique random forest to perform opinion mining on the data. Here,it observed that random forest algorithm gives more accurate result than other algorithms. It also shows the bar graph for particular team according to the opinion classes.

In the proposed system we have classifies the tweets in more accurate classes that is Strongly positive and strongly negative. In the future research, there is need study the face detection opinion mining of tweets.

VI. ACKNOWLEDGEMENT

I would like to thank my assigned guide Prof. N. G. Pardeshi whose interest, knowledge and guidance helped me a lot to complete the project work successfully. Also the very valuable help of Dr. D. B. Kshirsagar (HOD Computer Department.) who gives the facilities to explore and learn the concept in more details that helps to improve the result.

REFERENCES

[1] Arti, Kamanksha Prasad Dubey, Sanay Agrawal, —An Opinion Mining For Indian Premier League Using Machine Learning Techniques‖, 978-1-7281-1253-4/19/$31.00 © 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, —Machine Learning-Based Sentiment Analysis for Twitter Accounts‖, Mathematical and computational applications .,2018, 23, 11.

[5] Er. Hari K.C, —ONLINE SOCIALNETWORK ANALYSIS USING MACHINE LEARNING TECHNIQUES‖, International Journal of Advances in Engineering Scientifific Research, Vol.4, Issue 4, Jun-2017.

[6] Grigori Sidorov , Sabino Miranda-Jim´enez , Francisco Viveros-Jim´enez , Alexander Gelbukh , No´e Castro-S´anchez , Francisco Vel´asquez , Ismael D´ıaz-Rangel , Sergio Su´arez Guerra , Alejandro Trevi˜no , and Juan Gordon ,—Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets‖, Center for Computing Research, Instituto Polit´ecnico Nacional.

[7] Vidushi, 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 , 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, Alexander Gelbukh, ―Aspect extraction for opinion mining with a deep convolutional neural network‖, Temasek Laboratories, Nanyang Technological 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.