International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

An Opinion Mining for Indian Premier League Using Machine Learning Techniques

Rakshith N[1], Suraj B S[2], Chethana C[3]

PES College of Engineering, Mandya

Abstract: ***Social media has dramatically changed the way people express their opinion, appraisals or feelings towards entities or brand. Among many social media sites one of the free social service websites i. e. twitter, that permits users to publish their everyday life related events. As we know that twitter blog posts are being originated continually and Twitter having character short or limit to the Twitter posts (tweets) and also extremely compatible origins of continuous flow data for finding or detect opinion mining. Blog posts will reveal general or people emotions once taken in collection as an example throughout events like IPL 2016. Here our work presents, and provide the effectiveness of a machine learning model as positive or negative sentiment on tweets. The data collection of tweets and processing them by filtration best of the authorized IPL hashtags (#IPL 2016 and #IPL 9) which can be done through using of Twitter's API (Application Programming Interface) service. We analyze the performance of the 'Random Forest' against existing supervised machine learning algorithms with respect to its accuracy, specificity, sensitivity etc. here explanation of our paper is to performed opinion mining for the event like Indian Premier League 2016 effectively.***

Keywords: Social network, Opinion Mining, Machine Learning, Twitter, Stream Data Analysis

1. Introduction

Social networking sites or we can say media has become one among the foremost trendy communication tools for transference feelings and interacting with people within the on-line world. With the commencement of applications such as forums, micro-blogging and social networks, there came reviews, remarks, feedbacks, comments, critique and ratings generated by users. The user generated message can be about anything in this world for example products, movies, events, news, stars, politician, companies etc. This sheer volume of data can be mined to extract valuable information. People tweets about a particular product, review the product, or retweet product posts, it is very much possible that along with what they are replying, tweeting or retweeting to, they can articulate their opinion. This offers massive data for the purpose of sentiment analysis or opinion mining. Opinions are primitive to almost all human activities and are key influencers of our doings. The term opinion often reflect the states of individual's sentiments, understanding, desires, evaluation, emotions and speculation. Using polarity analysis, a proposal of public's feedbacks to various events often well-mined, which may give useful results and also additionally classifications [1]. One of the major challenges during this area of analysis is to enhance the sentiment prediction.

Opinions, speculations, emotions and evaluations typically reveal the states of different from class; they include opinionative narrow data expressed during a language which is compiled of subjective statement [2]. "Sentiment analysis task is to identify people's belief, assessments, thoughts, appraisals, sentiments, and feelings towards instances like services, outcomes, events, features and their framework." [3]. This paper provides information about public's perception towards an event IPL 2016.

2. Related Work

Tweet might contain both positive and negative sentiment concurrently so it is not easy to do sentiment analysis on twitter data as a [2]. in normal sentiment of a object or document for one individual issue then returns, Sentiment polarity. Pang et al. [4] in their work have explained machine learning techniques for sentiment polarity. Twitter can be considered a standard on which one can measure the performance of opinion classification on a multiplicity of objects, ranging from stock time series analysis [5] [6] [7] [8], health analysis [9] movie classification or box-office analysis [10] [11], crime prediction [12], sports [13] movie and election forecasting [14] [15]. Twitter posts can also be used to analyse sentiment towards products.

In Rishabh et al. [16] authors have used twitter to analyse sentiment towards a product 'iphone6s'. They used 'cluster-then-predict' method, in which a clustering of the tweets using K-mean algorithm was performed to improve the accuracy of the prediction. In the paper [17], authors Alexander Pak et al. categorise the feelings into superb, negative or impartial for that an automated corpus building method used, based on which a sentiment classifier was built.

In review there are various types of learning methods among them one of the supervise learning including Naïve Bayes, Support vector Machine (SVM) and CART which also can be used to classify sentiments. In [18], authors have used a sentiment classifier like Support Vector Machine, Maximum Entropy (MaxEnt), including Naïve Bayes which used many step by step process to visualize the twitter data. In [19], author Linhao Zhang has done a vast analysis on the effectiveness of the sentiment classification performed using tweeter data.

3. Methodology for Implementation

For sentiment analysis below depict the procedure to create or develop proposed model on twitter data. After obtaining the data and pre-processing it, a novel approach called 'Random Forest' is applied on the data for predicting the opinion for the event IPL 2016. Finally,

result's explore and encounter with alternative classification algorithms.

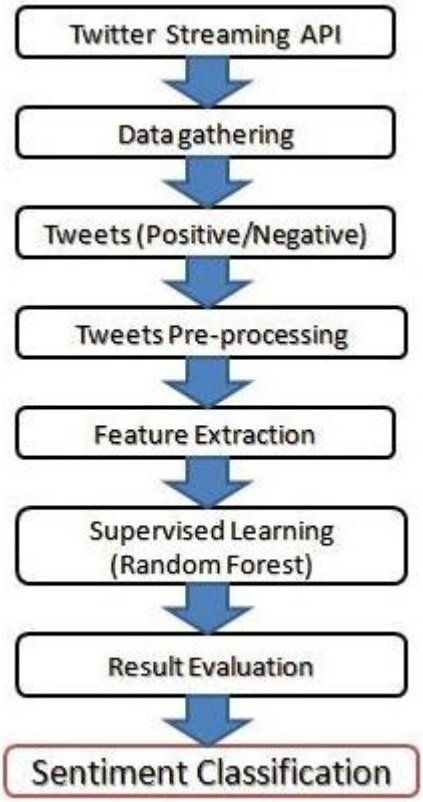


Figure 1: Flowchart of methodology

A. Data (Tweets) Compilation

To interact with data tweets Twitter's APIs has used which allows to access the data for opinion mining and containing the term 'IPL 2016'. Tweets were filtered and parsed which included the 'IPL 2016'and which showed some sentiment value towards it. Then the streaming Twitter's posts were stored to a structured or tabular file format which containing many instances and format of file commonly used as CSV.

The opinion rating of every tweet had been record into during distinct attributes. The opinion rating was contracted by those five people who are aware of sport event such as IPL and finding the average of rating which is given by sport experts. For mining the sentiment we have extracted a dataset of 3117 tweets, which having 1193 tweets as a opinion towards positive and 1924 tweets towards negative sentiment for 'IPL 2016'. 'R' is a language for statistical computing and graphics and it is useful in structural and statistical aspects. In 'R' studio containing many packages for sentiment mining and make number of things much easier. So that structured data format is brings to 'R' studio which support 'R' statistical programming language and having Graphical Interface setting environment [20].

B. Pre-processing Tweets

In a primary step towards finding a tweet's opinion and so as to get correct opinion classification, we like to filter the unnecessary term from the initial text of tweets that don't provide something to a tweet's emotions. Data cleanliness and preprocessing cannot solely shorten the classification task for the mathematical model however it additionally provides to greatly decrease process price during the training stage.

There are some steps for preprocessing to make sure that our procedure works exactly with tweets data.

1) In order to make all the words are either in upper-case or lower case so we have to alter the text.
2) Even punctuations marks like comma, period, semicolon, some special symbols e. t. c. can also issues so we have to eliminate some useless and meaningless words or symbols e. g "[at]IPL 2016", "#IPL 2016", "IPL 2016!" might be deal with 'IPL 2016'.
3) The several words that is "stopwords" like to be at, which, is, the, etc commonly used however only significant in an exceedingly plain text, which is not useful so eliminate them to minimize the dimension of data.
4) Deriving from-This process is intended by the importance to signify words with unlike compilation because the analogous word e. g plays, played; playing is simply changed to play.

C. Feature Extraction

Opting a helpful catalog of words as distinctive attributes of a plain sentence and eliminating a huge number or an outsized range of words that do not provide to the text's opinion is define termed as extract the features. "Bag-of-Words" technique is used for feature extraction [21]. This put up a document-term matrix or term document rectangular of elements, which creates a array of elements with rows compatible to instance (tweets) and columns compatible to the words in those instance or tweets. The data we tend to get is termed dispersed, implies that there are several zeroes in our array of matrix. The values within the array of components are the amount of times that word emerges in corresponding row. The significance words showing for the foremost part may be visualized in bar chart and tag (word) cloud (b) in below diagram by Using the document-term array of elements.

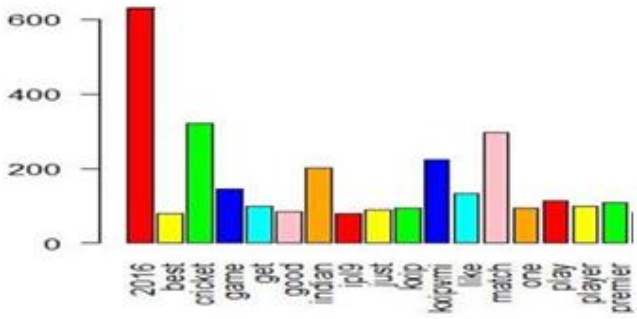


Figure 2 (a): Displaying bar chart of 'words frequency'

Figure 2 (b): Representing world mass

4. Performing Supervised Learning Using Random Forest

In order to classifying text analysis and assortment one of the foremost eminent and regularly method has been delineated below that is employed to train a learning model which is supported based on Twitter's tweets data. We have graphical representation of possible solutions to take the decision which is based on conditions. Produce many decision trees and combine them together in order to correct and stable classifying one of the widely used machine learning techniques that is called Random Forest. The training dataset divides into less than normal and normal components so as to acknowledge specimen which might be used for classification they have supervised process learning algorithmic program. The fact is then given kind of like a flow chart within the style of rational structure which will be merely understood with none any applied parametric statistics data.

Random forest is a versatile method which has capability of employed classification as well as regression problems. And it various trained model associate results used as to provide a additional correct classifier this type of method is known as Ensemble learning method this methodology introduced by (Breiman, 2001) [26],. A group of models have significantly developed performance than the single techniques or model. it's quite an easy rule however will turn out state-of-art performance in terms of prediction.

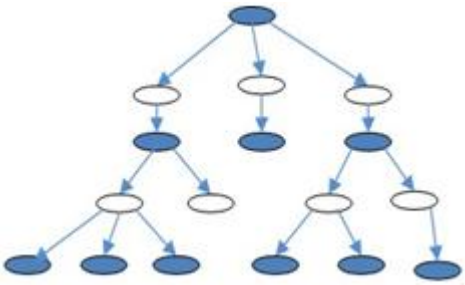


Figure 3: Decision Tree Structure

In order to provide the training to the dataset (tweets) the data is divided into two parts one is the training dataset containing 70% and other one is testing sets containing 30% of the dataset then among the various supervised method, Random Forest techniques has applied to training datasets.

5. Result Evaluation And Sentiment Classification

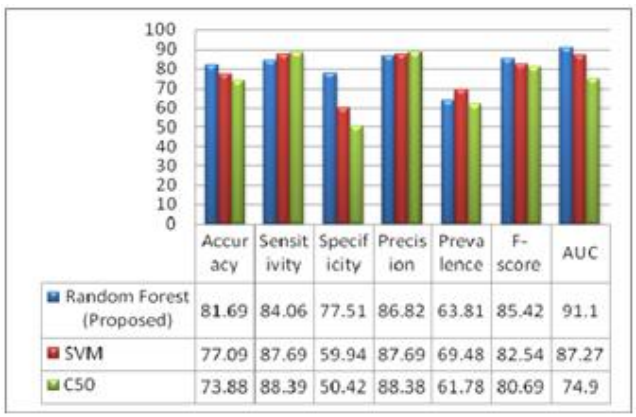
To classify the end result variable Random Forest techniques were used to the corresponding testing dataset. for each tweet, a two class value with '0' and '1' representing positive and negative opinions respectively shown as a output variable. A confusion matrix or error matrix is used to measure the performance of machine learning prediction. To find array of matrix we have written command in 'R' programming language. An error matrix contains details about true classification and predicted classifications. The execution of the experimental 'Random Forest' techniques was encounter with other existing system, like SVM, C50. and performance is evaluated based on the different parameters such as Accuracy of Prediction, which has depicted below Fig.5.

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Figure 4: Confusion Matrix

Define a confusion matrix is used for a two class classification problem, and we have set of documents which are to be tested. True positive (TP) - An actual class is positive instance and predicted class output is also positive. It means learning algorithm correctly classified.

- False positive (FP)-an actual class is negative instance but after learning algorithm comes positive. It means learning algorithm wrongly classified.
- True negative (TN)-an actual class is negative instance but after learning algorithm comes negative. It means learning algorithm correctly classified.
- False negative (FN)-a True class is positive instance but after learning algorithm comes negative. It means learning algorithm not correctly classified.



	Accuracy	Sensitivity	Specificity	Precision	Prevalence	F-score	AUC
Random Forest (Proposed)	81.69	84.06	77.51	86.82	63.81	85.42	91.1
SVM	77.09	87.69	59.94	87.69	69.48	82.54	87.27
C50	73.88	88.39	50.42	88.38	61.78	80.69	74.9

Figure 5: Chart layouts of Random Forest (Proposed) and other machine learning algorithms

6. Conclusion

In order 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.

In this paper illustrates the views of 'Random Forest' that has primary result on comprehensive accuracy of the interpretation or analysis. This proposed methodology has an accuracy of 81.69% for classification. So the comparison between totally different techniques and proposed method shows that proposed techniques is preferable in essential analysis measurable attributes of correctly classifying out of all examples (tweets), specificity, F-score and Area Under Curve.

In future work, author (s) will attempt to comprise the more accuracy of our build model with different ensemble techniques in order to classify displaying positive, strongly positive, neutral, negative or strongly negative opinions mining. For analyzing the classification problem might need more than two class classification problems.

References

[1] S. Asur and B. A. Huberman, "Predicting the future with social media, " in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 2010, pp.492-499.
[2] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data, "in Discovery science, 2010, pp.1-15.
[3] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis Lectures on Human Language Technologies 5, no.1 (2012): 1-167.
[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques, " in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp.79-86.
[5] Mittal, Anshul, and Arpit Goel: Stock prediction using twitter sentiment analysis. Stanford University, CS229 (2012).
[6] Zhang, Linhao: Sentiment analysis on Twitter with stock price and significant keyword correlation. Diss.2013. The University of Texas at Austin. (2013).
[7] Ohmura, M., Kakusho, K., & Okadome, T.: Stock Market Prediction by Regression Model with Social Moods. International Journal of Social, Behavioral, Educational, Economic and Management Engineering Vol: 8, No: 10, 2014.
[8] Vu, Tien-Thanh, et al.: An experiment in integrating sentiment features for tech stock prediction in twitter.: In: Workshop on Information Extraction and Entity Analytics on Social Media Data, 9 Dec 2012, Mumbai, The COLING 2012 Organizing Committee, 23-38. (2012)
[9] Mark E. Larsen, Tjeerd W. Boonstra, Philip J. Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. "We Feel: Mapping Emotion on Twitter" IEEE Journal of Biomedical and health informatics, VOL.19, 4, July 2015
[10] Lu, Yafeng, et al.: Integrating predictive analytics and social media. In: Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on. IEEE. Paris, France (2014).
[11] Pak, Alexander, and Patrick Paroubek: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC. Vol.10.2010.
[12] Wang, Xiaofeng, Matthew S. Gerber, and Donald E. Brown: Automatic crime prediction using events extracted from twitter posts. In: Social Computing, Behavioral-Cultural Modeling and Prediction. Springer Berlin Heidelberg, 231-238. (2012).
[13] Peiman Barnaghi, Parsa Ghaffari and John G. Breslin, "Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets" Conference ACM SIGKDD'15, Sydney, Australia, Workshop on Large-Scale Sports Analytics (LSSA), August 10-13, 2015
[14] Kagan, Vadim, Andrew Stevens, and V. S. Subrahmanian: Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election. In: IEEE Intelligent Systems 1 (2015): 2-5.
[15] Gayo-Avello, D.: "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" - A Balanced Survey on Election Prediction using Twitter Data. arXiv preprint arXiv: 1204.6441. University of Oviedo, Spain. (2012).
[16] Rishabh Soni and K. James Mathai. " An Innovative 'Cluster-then-Predict' Approach for Improved Sentiment Prediction" International Conference on Advanced Computing and Communication Technologies (ICACCT) proceedings "Springer Science + Business Media Singapore Pte. Ltd". (Nov 28, 2015).
[17] Pak, Alexander, and Patrick Paroubek: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC. Vol.10.2010.
[18] Selmer, O., & Brevik, M.: Classification and Visualisation of Twitter Sentiment Data. Master's Thesis, NTNU-Trondheim. (2013)
[19] Zhang, Linhao: Sentiment analysis on Twitter with stock price and significant keyword correlation. Diss.2013. The University of Texas at Austin. (2013).
[20] The R project for statistical computing, https://www.r-project.org/
[21] "Bag-of-Words" feature extraction technique, https://en.wikipedia.org/wiki/Bag-of-words_model