International Journal on Recent and Innovation Trends in Computing and Communication
Volume: 4 Issue: 6

ISSN: 2321-8169
172 - 175

Comparing Classification and Regression Tree and Support Vector Machine for Analyzing Sentiments for IPL 2016

1 Arti , 2 Sanjay Agrawal
1 PG Scholar, Department of Computer Engineering and Applications, NITTTR Bhopal,
Bhopal, Madhya Pradesh, India
2 Professor, Dean R&D,Department of Computer Engineering and Applications, NITTTR Bhopal,
Bhopal, Madhya Pradesh, India

Abstract:-Social media is incredibly popular method for expressing opinions and interacting with other people in the online world. Twitter is one of the most frequent online social media and micro blogging services. It enables users to communicate with others and get updates on topics and events that interest them. Tweets can reflect public attitude when taken in aggregation, for example during events such as IPL 2016.Machine learning makes sentiment analysis more effective. In this paper, we examine the evaluation of machine learning algorithms (CART and SVM) in R to find the public opinions about event IPL 2016.

Keywords:*-Twitter, Sentiment Analysis, Opinion Mining, Sentiment Classification, Natural language Processing*

*****

1. INTRODUCTION

Social media is extremely well-liked communication tools for sharing opinions and everyday life-related actions. Millions of posts are found on social networking sites like twitter1, instagram2, facebook3 which could be used in advertising and study of the user information of social network is one of the current vogue of the times. The informal nature of Twitter leads to a lot of opinions being posted and this has made twitter a valuable research resource for Sentiment Analysis. Tweets (Twitter posts) has maximum of 140 characters. We use data collection from twitter which is a kind of content. The content of messages alters from individual to social aspect. Bulk of the sentiment that is available on the internet is of an unstructured format. It is not easy for computers to process it and educe meaningful information from it. NLP tools and techniques are used to transform this raw data into a format that can be processed efficiently by a computer, and end results could be properly visualized to gain insights from it.

Sentiment Analysis alludes to the field of Natural Language Processing (NLP).Sentiment analysis also know as opinion mining. It is challenging natural language processing or text mining problem. Sentiment analysis means the mining of an opinion's overall polarization and potency towards a specific subject. In this paper we propose an analysis of collective sentiments related to IPL 2016.

There are basically three paradigms of learning (shown in figure 1). In this paper, the result of sentiment analysis is carried out to get public perception for 'IPL 2016' using supervised learning in the form of Classification and Regression Trees (CART) and Support Vector Machine (SVM), to effectively classify the data as belonging to positive or negative sentiment.

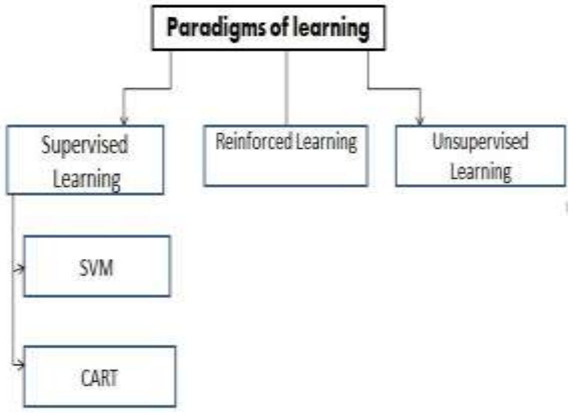


Figure 1. Types of learning

2. RELATED WORK

Opinion mining returns the overall sentiment to classify movie reviews into two classes positive and negative. Pang et al [1] provided a various features and machine learning algorithms such as Naïve Bayes , maximum entropy classification and support vector machine.

Juliano et al.[2] used collective classification techniques which use link information to infer sentiments of users who have not tweeted their political opinion. They used a graph representation in which nodes signify users and edges signify their association in social network.

Rishabh et al [3] proposed a hybrid approach which aggregates unsupervised learning in the pattern of k- means clustering to cluster the tweets and then perform some supervised learning methods such as support vector machine and decision trees (CART).The model given by author is known as cluster-then-predict model. He showed the sentiment the users have towards the product 'iPhone 6s' using R language.

Peiman et al [4] used a well-know supervised machine learning methods for text analysis and sentiment polarity, called Logistic Regression Classification (LRC) using WEKA. They studied the attitude of public i.e. positive or negative during FIFA World cup 2014.They tried to perform the opinion polarity detection using their trained model to find links between tweets and events that occurred during the World Cup tournament .The author collected 4162 tweets and labeled those tweets manually("positive" and "negative").

Sinha et al [5] proposed tweets as social media output to find correlation with National Football League (NFL) games. They forecast game and gambling outcome using logistic regression classifier. They collected tweets as a weekly, pregame and postgame and tried to find standard coefficient based on all games.

Singh[6] attempted to use Data Envelopment Analysis(DEA) to benchmark and compute technical efficiency of cricket teams in Indian Premier League 2009.The input used in this session is used by the teams moved towards by the total expenses and output is measured by the points awarded, net run rate, revenues and profits.

Vasuki et al [7] used twitter a great source of information to monitor public's feeling towards particular brands or events for decision making using sentiment analysis. The author attempted different pre-processing steps before feeding the text to the map reduce algorithm. He proposed map reduce to get precise decision about product and to achieved higher precision when compared to analogous technique.

Due to varied nature of textual information available on twitter, various analyses like public opinion on movie reviews[1],political issues[2],product "iphone 6s",[3] event "FIFA World Cup 2014",[4] NFL game[5],technical efficiency of cricket team IPL [6],brands or events for decision making through Map Reduce.

In this research, authors have used sentiment analysis techniques for finding out people's insight about the event.

3. PROPOSED WORK

Figure 2 shows the steps taken to build a model for opinion mining. The trained model is used for polarity detection on Indian Premier League 2016.

Author proposes a supervised learning for predicting the opinions of public about Indian Premier league 2016.There are various supervised learning methods for building the model, but author found SVM is foremost accepted for the problem because it provides best trade-off in accuracy, recall, error rate, Area under curve (AUC) when compared to CART.

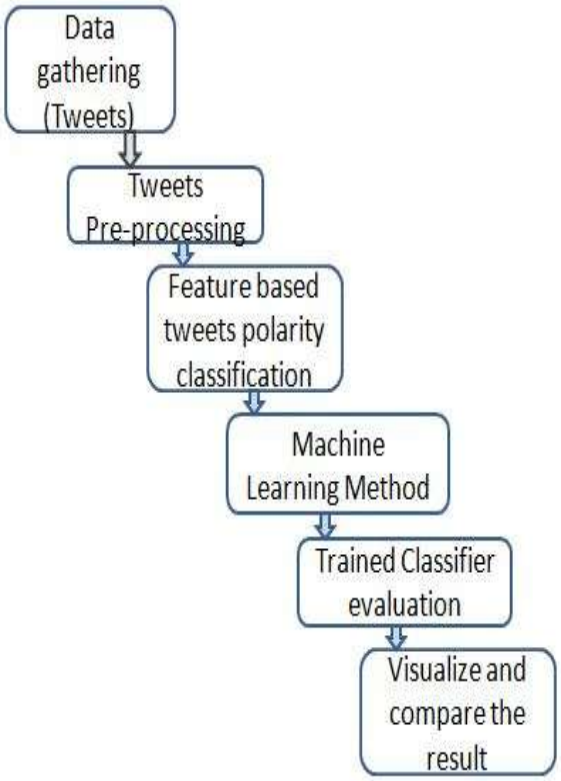

Data gathering (Tweets)
Tweets Pre-processing
Feature based tweets polarity classification
Machine Learning Method
Trained Classifier evaluation
Visualize and compare the result

Figure 2 Proposed Methodology

3.1 Collecting Data

We used Twitter (www.twitter.com) as a origin of social media messages ("tweets").Twitter provides Streaming APIs to interact with their service. Twitter's streaming API is usually true for live events with a world-wide coverage. The IPL event starts from 9th April 2016 to 29th May 2016. The tweets had been collected before the event, and between the event. Python's API named tweepy [8] is most interesting and straightforward that has been used to implement streaming API of Twitter. In this way the incoming twitter posts a can be easily collected as it provides libraries. The incoming twitter posts were reserved in CSV (Comma Separated Values) file format in real-time by importing Python's CSV library functions. This consisted of 3118 tweets manually marked "negative" or "positive" [3].The author was collecting tweets during the indian premier league 2016 and processing them by filtering some of the official Indian Premier League hash tags (e.g "#ipl2016", "#ipl").

3.2 Tweets Pre-processing

As a first step towards finding a tweet's post is completed now we need to clean some noise and worthless symbols from the original text of tweets that do not contribute to a tweet's opinion. Fully understanding text is very difficult but bag of words [9] provide it simple. As it counts the number of times each word appear. One part of cleaning the text is cleaning up irregularities, as text data often has many inconsistencies that will cause algorithm trouble e.g. @IPL, ipl,--ipl-- will be considered three terms if cleaning the text will be undone. Stemming is an essential part of pre-processing because we do not need to draw distinction among words for example play, plays, played, playing could all be represented by a common stem "play".

3.3 Feature based tweets polarity classification

Labeling the persuasive text grouping it overall into a positive ,neutral and negative class is known as sentiment polarity ranking. The neutral label is used for more purpose items that have a lack of emotion in the text, or when person is disinterested or unbiased.

3.3.1 Feature extraction

Determining an effective list of words as quality of a text and discarding a large number of words that do not contribute to the text's opinion is explained as feature extraction. It helps us to filter out noise from the text and get more precise opinion for tweets. In this paper, we use N-grams feature for opinion classification for Indian Premier League 2016.

3.3.1.1 N-grams Feature

N-grams are defined as taking a set of sequential words in a text; for example n= 1,2,3 or 4 character arrangements is referred to as unigrams, bigrams, trigrams or tetra grams respectively.

Related work based on unigrams shows its importance on classification working. Pang et al [1] show that unigrams contribute better performance on movie reviews for opinion polarity.

3.4 Machine Learning Methods

Machine learning is a subfield of computer science that evolved artificial intelligence which emphasizes on producing models that have the property to learn from data. In 1959, Arthur Samuels's general definition relate it well:

"Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed" [9].

The proposed approach of classification utilize supervised learning. In supervised learning, it is assumed that the correct "target" output values are known for each input patterns. Figure 3 represents supervised learning where X is input, Y is actual output,(D-Y) is error signals and D is desired output.

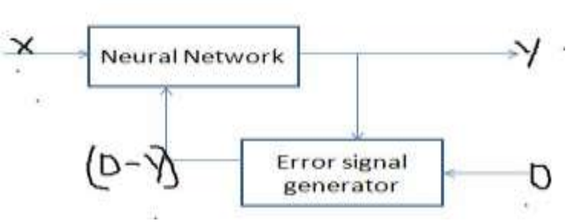


Figure 3 Supervised Learning

Our aim in this work was to examine the sentiments of public about ipl 2016(i.e positive or negative sentiments) and also to find out which machine learning algorithm gives best accuracy either SVM or CART.

To implement these machine learning algorithms on our tweets we used R language.

3.4.1 Support Vector Machines

The support vector (SVMs) is one of the most refined algorithm. It has been shown to be highly valuable at traditional text categorization. The support vector machine is classified as a non-probabilistic binary linear classifier. It works by plotting the training data in multidimensional space. Then it tries to separate the classes with a hyperplane. If the classes are not immediately linearly separable in the multidimensional space the algorithm will add a new dimension in an attempt to further split the classes. It will continue this process until it is able to distinguish the training data into its two separate classes using a hyperplane. Figure 4 basically represents how it splits the data.

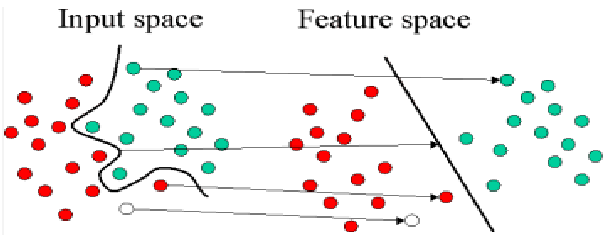


Figure 4 SVM basic operation

3.4.2 Classification and Regression Tree

Classification and Regression Tree(CART) are important aspects of Machine Learning. CART is a popular decision tree algorithm. It was first published by Breiman in 1984.The CART algorithm grows binary trees and continues splitting as long as new splits can be found that increase clarity. As shown in figure 5, inside a complex tree, there are many simple subtrees each of which represents a different trade-off.

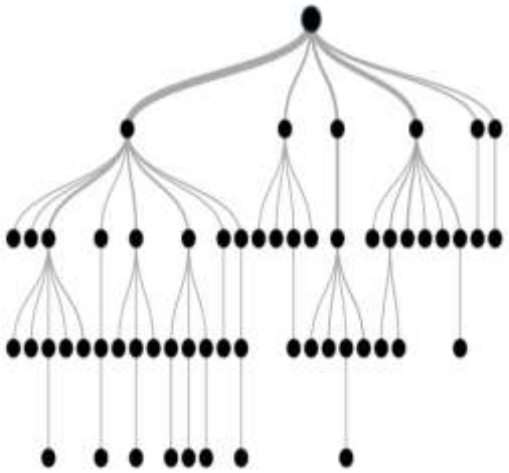
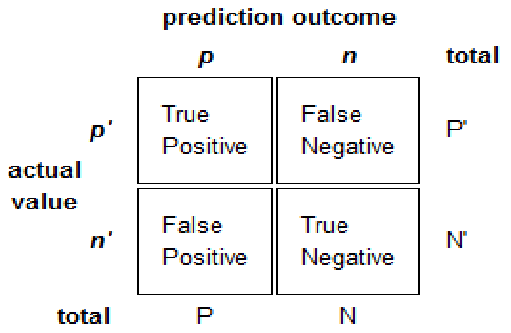


Figure 5 Decision Tree

4. TRAINED CLASSIFIER EVALUATION

The tweet polarity classifier is trained based on N-grams feature (N=2) using R as text mining and machine learning classifiers.

In this phase with the help of confusion matrix we evaluate that how CART and SVM classifiers affects the accuracy, recall, f-measure, AUC from each other.



		prediction outcome		
		p	n	total
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
	total	P	N	

Figure 6 Confusion Matrix

The accuracy, recall, error rate and AUC of the models has been taken out by Confusion matrix.

5. VISUALIZE AND COMPARE THE RESULT

Techniques	Parameters			
	Accuracy (%)	Recall	Error Rate	AUC
SVM	82.33%	0.8433	0.176659	0.9359644
CART	71.73%	0.7035	0.282655	0.7628890

The table here shows that SVM approach has better accuracy, recall and AUC and less error-rate.Error rate is also known as misclassification rate.

6. CONCLUSION

In this paper, a sentiment classification model is trained based on twitter's posts using text features. We extracted the sentiment polarity for IPL 2016 using our trained model. The experimental results manifest the positive and negative reaction of people towards IPL. The architecture of this model is scalable, so it can entertain ample amount of twitter text data.

In future work, author(s) will try to include sentiment from emotions, and also will try to take most used languages in tweets other than English tweets.

REFERENCES

[1] 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-2002

[2] J.Rabelo, R.Prudencio and F. Barros,"Collective Classification for Sentiment Analysis in Social Networks" in IEEE 24th International Conference on tools with Artificial Intelligence-2012

[3] R.Soni and J.Mathai "Improved Twitter Sentiment Prediction through 'Cluster-then-Predict Model'" in International journal of Computer Science and Network,-2015

[4] P.Barnaghi,P.Ghaffari and J.Breslin "Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets"in ACM Conference on Workshop on Large-Scale Sports Analytics- 2015

[5] S.Sinha,C.Dyer,K.Gimpel and N.Smith"Predicting the NFL using Twitter" arXiv preprint arXiv:1310.6998, 2013.

[6] S.Singh"Efficiency Evaluation of Teams in IPL"in Recent Researchers in Applied Mathematics and Informatics

[7] M.Vasuki,J.Arthi and K.Kayalvizhi "Decision Making using Sentiment Analysis from Twitter" in International journal of Innovative Research in Computer and Communication Engineering"-2014

[8] Python's API for Twitter: http://www.tweepy.org/

[9] Bag-of-Words feature extraction technique: https://en.wikipedia.org/wiki/Bag-of-words_model

[10] Samuel, Arthur L. "Some studies in machine learning using the game of checkers." *IBM Journal of research and development* 44-2000