PREDICTING THE CRICKET MATCH OUTCOME USING CROWD OPINIONS ON SOCIAL NETWORKS: A COMPARATIVE STUDY OF MACHINE LEARNING METHODS

Raza Ul Mustafa[1], M. Saqib Nawaz[2*], M. Ikram Ullah Lali[3], Tehseen Zia[4], Waqar Mehmood[5]

[1] Department of Computer Science, COMSATS Institute of Information Technology, 57000 Sahiwal Campus, Pakistan
[2] Department of Information Science, School of Mathematical Sciences, Peking University, 100871 Beijing, China
[3] Department of Computer Science and Information Technology, University of Sargodha, 40100 Sargodha, Pakistan
[4] Department of Computer Science, COMSATS Institute of Information Technology, 44000 Islamabad, Pakistan
[5] Department of Computer Science, COMSATS Institute of Information Technology, 47040 Wah Campus, Pakistan.

[*]Email: msaqibnawaz@pku.edu.cn

ABSTRACT

Social media has become a platform of first choice where one can express his/her feelings with freedom. The sports and matches being played are also discussed on social media such as Twitter. In this article, efforts are made to investigate the feasibility of using collective knowledge obtained from microposts posted on Twitter to predict the winner of a Cricket match. For predictions, we use three different methods that depend on the total number of tweets before the game for each team, fans sentiments toward each team and fans score predictions on Twitter. By combining these three methods, we classify winning team prediction in a Cricket game before the start of game. Our results are promising enough to be used for winning team forecast. Furthermore, the effectiveness of supervised learning algorithms is evaluated where Support Vector Machine (SVM) has shown advantage over other classifiers.

Keywords: Pattern Recognition, Cricket, Twitter, Sentimental Analysis, Opinion Mining

1.0 INTRODUCTION

Starting from instant chat, and then with email, newsgroups, forums and blogs for discussion, people have engaged in online social interactions since the beginning of the Internet. Recently, social media has become ubiquitous and important for social networking and content sharing. Social media are helping people to create their own contents, share it, participate in activities and live events, follow breaking news and keep up with friends and families. Social media is changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry [17]. Today, Facebook boasts 1.3 billion users, with 82% of users are from outside the United States and Canada. Twitter has 270 million active users and 500 million tweets are sent per day. Furthermore, each day, more than four billion videos are viewed on YouTube and 60 million photos are uploaded on Instagram [13].

Twitter is a popular platform for opinion and information sharing, and this platform is mostly used before, during and after live events. Twitter effectively takes part in any mega event happening around the world. Data obtained from Twitter has been effectively utilized in the prediction and explanation of various real-world phenomena, such as spreading of infectious diseases [14], elections [4], stock market prediction [9], opinion polls [5] and sport results [6, 15, 18, 22]. On average, Twitter posts contain meaningful information that can be exploited with the help of statistical methods. Therefore, Twitter offers a way to exploit the *wisdom of crowds* [10] concept to make better predictions of the real-world events.

The *wisdom of crowds* is a notion that reflects the cumulative opinion of a diverse and large group of individuals. This information is often proved to be more useful as compared to expert judgment. Harnessing this notion for making predictions has been the subject of many studies. A number of researchers have used social media as a meaningful source of information for sports matches outcome prediction. Cricket is the second most popular sport in the world after Soccer with two to three billion fans [21]. In 2014, Indian Premier League (IPL) was followed by 1.5

million people on Twitter and five million opinions were posted [3], and during Cricket World Cup (CWC) 2015, the craze of fans were even greater. The complex rules used in Cricket match and various natural parameters (such as batsman, bowlers and fielders skills and performances, advantage of home game, match format (Day/Night), the venue of match and weather conditions) that affects the outcome of a Cricket match present a great challenges for accurate prediction of the match results. Furthermore, growing interest in strategic planning for ensuring victory before match has motivated prediction of future games. In formulating successful strategies, one of the key problems that need to be solved is to predict the outcome of a match.

Opinion mining and sentiment analysis techniques are used to detect and extract subjective information in text documents [33]. Using sentiment analysis, we can find the overall contextual polarity about any topic provided by the author. The challenging task in opinion mining is sentiment classification which is done by guessing the opinion about anything i.e. book, movie, product, issues regarding politics and religion etc. These opinions can be in the form of sentence, document or feature, and the task is to label them as positive, negative or neutral. In this work, we have applied opinion mining techniques on tweets related to Cricket matches. Our goal is to predict the winner of Cricket match using microposts posted on Twitter. For this, we have collected tweets of 109 matches played during IPL2014 and CWC2015. These microposts were picked from official Twitter page of Cricket teams and from popular Cricket websites. It is found that these microposts were mostly written in English. From the extracted tweets, three features are extracted that include tweets' volume, aggregated fans' sentiments and score predictions. Finally, Support Vector Machine (SVM), Naïve Bayes (NB) and Logistic Regression (LR) are used for training and evaluation purpose.

The rest of the article is organized as follows. Related work is discussed in Section 2. The methodology that we used for predicting the outcome of a Cricket match is introduced and explained in Section 3. In Section 4, we evaluate the performance of our devised method on 49 matches played during CWC2015 and 60 matches played during IPL2014. We choose CWC2015 as it is a mega event that is played between top teams every four years and IPL2014 due to the fact that the participating teams within a national league are much more intertwined. Finally, in Section 5, this article is concluded with some remarks.

2.0 RELATED WORK

Our work to predict the outcome of sports game is certainly not the first one. A limited but increasing number of academic researchers have performed some work using the social media data for prediction of American Football League (NFL) [18], English Premier League (EPL) [1, 8] and Soccer (World Cup) match results [15]. Rue and Salvesen [8] built a generalized linear model to estimate the skills of participating teams in EPL and to predict the outcome of Soccer games. However, due to many parameters, model prediction rate was low and out of 48 EPL matches only 15 matches were correctly predicted [8]. Joseph *et al.* [1] predicted the outcomes of Soccer matches in EPL played by a single team - Tottenham Hotspur. Their model depends on the presence of particular players and therefore, their model is not scalable.

Hong and Skiena [22] presented a study on the relationship between NFL betting line and public opinion expressed in blogs and Twitter microposts. They used sentiment analysis on various data sources, including Twitter microposts for predicting point spread of NFL games. Their prediction is based only on the positive and negative sentiments. UzZaman *et al.* [15] proposed a prediction system *TwitterPaul* for Soccer results prediction during-World Cup 2010. In *TwitterPaul,* context-free grammar is used for parsing the predictions made by users in microposts on Twitter in English. However, limited numbers of English speaking nations participated in the World Cup 2010. Furthermore, they also ignored negation within the microposts. They used the metric of *Root Mean Square Error (RMSE)* and their system does not beat the betting probabilities. Sinha *et al.* [18] used Twitter volume and the feature vectors of unigrams for predicting NFL matches. The information obtained from Twitter microposts only enhanced the accuracy to predict winner *with the spread* and the *over/under* over statistical knowledge, but not the prediction of the winner. In their work, Sinha *et al.* [18] also proved that statistical information can also be helpful in improving prediction based on collective knowledge. Godin *et al.* [6] developed four different methods for predicting the outcome of Soccer games in EPL. They have combined the four developed prediction methods with three

approaches to improve the accuracy of prediction, and their results showed that the combination of both statistical and collective knowledge can beat expert and bookmakers predictions.
Cricket is an entirely different game than Soccer. The outcome of a Cricket match depends on many factors such as: team compositions, pitch conditions and weather. Till now, some work has been done in predicting Cricket matches using machine learning and artificial intelligence. Duckworth and Lewis [34] introduced the D-L (Duckworth-Lewis) method for the adjustment of scores with balls in relation to the time lost (due to match interruption such as rain, poor visibility etc.) during match. International Cricket Council (ICC) accepted this proposal for target resetting in matches where time is lost due to match interruptions. Bandulasiri [35] investigated the effect of D-L method to predict the winner and concluded that the method does not have sufficient amount of information for match outcome prediction. Silva and Swartz [36] found that winning the toss provides no competitive advantage to a team but playing on one's home field does have some advantage. Allsopp and Clarke [37] established that home teams generally enjoy a significant advantage in test matches. They also concluded that generally, teams gain no winning advantage by winning the toss.

Choudhary *et al.* [38] used artificial neural networks to predict the outcome of Cricket tournaments. For training, they used match results played by the teams in the past 10 years. It was assumed that the players have not changed much over the past 10 years. Kaluarachchi and Varde [29] used association rules and Bayesian classifiers to predict how factors such as advantage of home game, day/day-night game, winning the toss and batting first affect the outcome of One-Day International (ODI) Cricket matches. Raj and Padma [11] analyzed the data of Indian Cricket team ODI matches and mine association rules from the following set of features: toss, home/away game, batting first or second and game result. However, both approaches used ~~a~~ limited subset ~~of~~ features for analyzing the factors that contribute to victory. Furthermore, both methods do not address score prediction and the overall game progress.
Bailey and Clarke [12] used historical match data for predicting the overall innings score with the help of linear regression. Their prediction model is updated when data of a match in progress streams in. Swartz *et al.* [19] used Markov Chain Monte Carlo methods to simulate ball by ball outcome of a match with a Bayesian Latent variable model. On the basis of features of current batsman, bowler and game situation, the outcome of the next ball is estimated. Both [12] and [19] have developed match simulators for ODI Cricket, but their models depend on games that are played over 10 years ago. In last 5 years, many rules of Cricket especially in ODI Cricket have changed.

Sankaranarayanan *et al.* [21] build a prediction system that analyzes historical Cricket match data and the instantaneous state of a match to predict game progression and the outcome of ODI match. They model the game using a subset of match parameters with linear regression and nearest-neighbour clustering algorithms. They used ridge regression and attribute bagging algorithms on the features for incremental prediction of the runs scored in the innings. Our approach is different from all previous approaches as we focused on extracting meaningful features from Twitter data for match outcome prediction. Furthermore, we also gave a comparative analysis of classifiers (SVM, NB and LR) to select the appropriate classifier.

3.0 METHODS FOR THE PREDICTIONS

The workflow of this research is shown in Fig. 1 and mainly composed of training and testing phases. In training phase, three main tasks are performed sequentially including tweets' collection, feature representation and classifier training. The testing phase is composed of four tasks: tweets collection, feature representation, hypothesis' prediction and evaluation. The first two tasks (i.e. tweets collection and feature representation) are shared between training and testing phase. The tasks are briefly elaborated below:

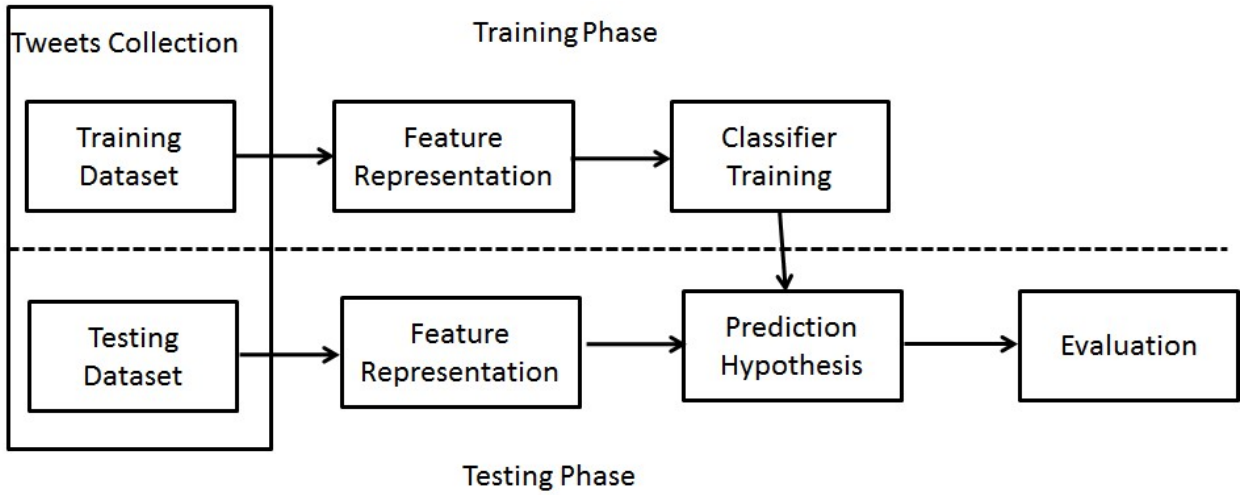

Fig. 1. Block diagram for the proposed methodology

3.1 Tweets Collection

The pre-match tweets of 60 matches played during IPL2014 and 59 matches of CWC2015 are collected from April 13, 2014 to June 1, 2015. For every team in the CWC2015 and IPL2014, we had a list of (nick) names along with the most commonly used hashtags for that team. Mostly tweets were picked from Twitter pages for CWC2015 and IPL2014, from popular Cricket websites including *cricinfo.com*, *cricbuzz.com* and official pages of team playing in these tournaments. Both stream and search APIs [20] are used for tweets extraction. For CWC2015, tweets were collected using Twitter API on daily basis during matches.

CWC2015 was played in Australia and New Zealand and 12 teams participated in that event. So majority of the games were away games for all teams. Therefore, we did not include home or away team feature for both CWC2015 and IPL2014. In total, four million tweets were posted on Twitter by users and the overall statistics for each tournament is shown in Table 1.

Table. 1. Twitter statistics for CWC 2015 and IPL 2014

	CWC2015	**IPL2014**
Total # of sites	15	11
Total # of tweets	2.3 M	1.7 M
Average # of tweets per match	46.9 K	28.3 K

3.2 Feature Representation

In this phase, three different aspects of twitter data is considered including twitter volume (TV), aggregated fans' sentiments (FS) and score predictions (SP). The total volume of microposts is a good indicator for team ranking [6, 18], so we extract the TV feature for each team as follows:

$$TV_i^m = \frac{\#\ of\ t_i^m}{n^m} \qquad (1)$$

Where TV_i^m indicates the volume of pre-match tweets of i^{th} team for m^{th}match, n^m is the total number of tweets for m^{th}match and $\#\ of\ t_i^m$ is the count of tweets of i^{th} team for m^{th}match.

The sentiment of Twitter data has shown to be an effective feature for future prediction in domains such as flu trends [39], elections [4] and in stock market [9]. To investigate the utility of sentiment score for predicting outcome of a match, linguistic features are selected for finding sentiment of tweets. Selection of these features is made on the basis of TF-IDF score (i.e. term frequency inverse document frequency) per class bases [40]. The selected linguistic features are shown in Table 2. These features are used to differentiate tweets into positive and negative classes (note: what?). By removing irrelevant tweets (those with only hashtags and those without any team name), the pre-match sentiment score of a team is calculated as:

$$FS_i^m = \frac{\#\,of{+}VE_i^m}{n^m} \qquad (2)$$

Where FS_i^m indicates sentiment score of i^{th} team for m^{th} match, n^m is total number of pre-match tweets for m^{th} match and $\#\,of + VE(t_i^m)$ is a count of positive tweets of i^{th} team for m^{th}match.

Table 2. Linguistic features for sentiment analysis

Positive	Negative
win, victory, cruise, top, good, strength, brilliant, phenomenal, wow, patchy, well done, excellent, nice, majestic, terrific, deserved, spectacular, high-class, awesome, wonderful, best, delight, comeback, massive, perfect, favorite, outstanding, great, dominate, flatten, ovation, convincing, blinder, emphatic, magic, promising, flawless, classic, likely, influential, all the best, strong, celebration, impressive, well played, cheer, out played, splendid, mighty, superb, nailed, triumph, fabulous, astounding, solid, sufficient, good luck, encouraging, matchless, sublime.	loss, defeat, suck, bad, crumble, terrible, unwatchable, awful, boring, stupid, worse, waste, poor, torment, unconvincing, hopeless, regret, thrash, disappointing, furious, alas, vulnerable, blow, worthless, unlikely, destroy, pointless, hard luck, unfortunate, dismantle, wreck, deprive, ruin, dismal, over, needless, mockery, unlikely, weak, all over, disaster, shabby, amateur, unimpressive, wasted, collapse, bad luck, flop, unexciting, typical, falling short, unachievable, unable, exposed, setback, unbelievable , trouble, crush, shaky.

Sample of tweets collected before a match (semifinal match between Australia and India) on the basis of linguistic features is shown in Fig. 2.

Fig. 2. Sample of positive and negative tweets for Australia team

The third method depends on finding the aggregate of score predicted by fans on Twitter. UzZaman *et al.* [15] showed that the result of aggregating fans predictions with the help of a context-free grammar can possibly beat the betting market. However, developing context-free grammar for such predictions is error prone, language dependent and time-consuming [6, 23, 24]. Therefore, we only use tweets that contained scores rather than written predictions. Note that Cricket fans generally do not predict the accurate score but give a limit such as "Australia will score around 280 and India will score 240." We divided the tweets into positive and negative classes for each team. This is done by defining the following regular expressions and rules:

- If in a micropost, only one team was refereed, we assume that specified team was the winning team. However, if the micropost contained words from a negative word dictionary that we built, then the mentioned team is assumed to be the losing team.
- In case both teams were specified in microposts, we looked for the following patterns: *\Team1 - Team2 score1 - score2* or *\Team1 score1 - score2 Team2*. In this case we assumed team on first position will win.
- For keywords such as *Team 1 will win* or *Team 2 will win*, we refer to such team as winning team.

For each team, we aggregated the predicted score for each match as follows:

$$SP_i^m = \frac{\sum_{j=1}^{l^m} s_{i_j}^m}{l^m} \qquad (3)$$

Where SP_i^m indicates the average predicted score of i^{th} team for m^{th}match, l^m is total number of tweets containing predicted scores for m^{th}match and $s_{i_j}^m$ is the predicted score of i^{th} team for m^{th} match in j^{th} tweet.

To evaluate the efficiency of this representation, three well-known classification algorithms are employed including SVM, NB and LR. Since each of these classifiers make different assumption about data, so comparative analysis between classifiers is an additional objective.

3.3 Classifier Training

A brief introduction of SVM, NB and LR is given below.

Support Vector Machine (SVM)

SVM method was primarily introduced by Joachims [32] for text categorization problem. Risk minimization (RM) principle is the basic idea of SVM. RM principle focuses on discovering a hypothesis to assure lowest true error. Unless the true target is known to learner, it is very difficult to make direct estimation of true error. However, using training error and complexity of hypothesis, the true error can be bounded. The main focus of SVM is to maximize reduction of true error of resultant hypothesis by controlling VC dimension efficiently where VC dimension (Vapnik Chervonenkis dimension) [27] is used to measure the capacity of the hypothesis space.

Geometrically, SVM can be explained with a binary classification problem as shown in Fig. 3. From Fig. 3, it is clear that we can select different separating hyper planes as a decision boundary. SVM selects the one that maximizes distance (also called as margin) with respect to instances laying on the boundary (called as support vectors) positive class. The bold line hyper plane is best among the others shown in dotted lines.

The problem of finding maximum margin is mathematically outlined as:

$$minimize_{w,b} < w.w >$$

$$Subject\ to\ y_i(< w.x_i > + b) \geq 1 \qquad i = 1,\ldots l$$

Where l represents number of training examples, x_i is the input vector, y_i is the desired output. Due to computational convenience, the problem is reformulated as:

$$L(w, b, \alpha) = \frac{1}{2} < w.w > - \sum_{i=1}^{l} \alpha_i [y_i (< w_i . x_i > + b) - 1] \qquad (4)$$

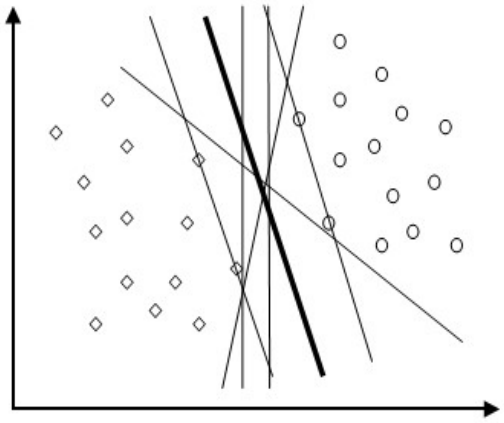

Fig. 3. A prototypical problem for learning SVM

Where $\alpha_i \geq 0$ are langrage multipliers. The langrage formulation in *Equation 4* is often termed as primal formulation. By differentiating *Equation 4* with respect to *w, b* and substituting their values in *Equation 4*, we can formulate the problem in so called dual form:

$$L(w, b, \alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j < x_i x_j > \quad (5)$$

When data instances are not linearly separable, then some sophisticated mathematical transformations are also applied (called as kernel trick) on the data prior to learning decision boundary. The main goal of transformation is to transform data instances to higher dimensions features space X. The instance x_i in new feature space is referred as Φx_i. An essential computational operation of the dual formulation is to compute the dot product between data instances that can be computed in higher dimensional space as $\Phi(x_i).\Phi(x_j)$.

The example is a binary classification problem where circles represent the instances of negative class and diamonds are the instances. Despite using this computationally inefficient approach, kernel functions $k(x_i, x_j)$ provide us a convenient way for this computation:

$$\Phi(x_i).\Phi(x_j) = K x_i, x_j \qquad (6)$$

On the basis of kernel function, SVM has three popular variants: SVM with linear or no kernel, polynomial kernel and radial basis kernel.

Naïve Bayes (NB)

In NB classifier, text categorization is viewed as estimating posterior probabilities of categories given documents- i.e. $P(c_i|d_j)$; the probability that j^{th} document (as re-presented with a weight vector $d_j = < q1_j, q2_j, \ldots q|T|_j >$ where qK_j is the weight of k^{th} feature in j^{th} document) belongs to class c_i [28]. These posterior probabilities are estimated by using Bayes theorem as:

$$P(c_i|d_j) = \frac{P(d_j|c_i)P(c_i)}{P(d_j)} \quad (7)$$

Where $P(c_i)$ is the prior probability that represents the probability of selecting an arbitrary (random) document that is part of class c_i, $P(d_j)$ is the probability (randomly) that an arbitrarily chosen document has weight vector d_j and $P(d_j|c_i)$ is the probability that the document d_j belongs to class c_i [2]. However, the guess of $(d_j|c_i)$ is intractable due to the problem of estimating $(d_j|c_i)$ involves very high dimensional vector d_j. To make computation tractable, it is assumed that document vector coordinates are conditionally independent of each other. By following the assumption, the term $(d_j|c_i)$ can be estimated as:

$$P(d_j|c_i) = \prod_{k=1}^{|T|} P(W_{kj}|c_i) \quad (8)$$

Logistic Regression (LR)

LR uses statistical analysis to predict an event that is based on known factors. LR can make predictions about whether a customer will buy a product based on age, gender, geography and other demographic data [27]. LR is another generalized linear model (GLM) procedure using the same basic formula, but instead of the continuous *Y*, it is regressing for the probability of a categorical outcome. In simplest form, this means that we are considering just one outcome variable and two states of that variable- either 0 or 1.The equation for the probability of *Y = 1* looks like this:

$$x = \frac{1}{1+e^{-(b_o+\sum(b_i X_i))}} \quad (9)$$

Independent variables X_i can be continuous or binary. The regression coefficients b_i can be exponentiated in order to give the change in odds of *Y* per change in X_i, i.e. $Odds = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=0)}$ and $\Delta_{Odds} = e^{bi}$. Δ_{Odds} is called the odds ratio, $\frac{\text{Odds}(X_i+1)}{\text{Odss}\,(X_i)}$. In simple words, we can say that the odds of *Y=1* increase by a factor of e^{bi} per unit change in X_i.

3.4 Prediction Hypothesis

The outcome of training a classifier is a hypothesis that can be used for predictions. By training SVM, NB and LR, we obtained three hypotheses for predicting the outcome of a match. In order to assess the effectiveness of these hypotheses, we performed evaluation task as described below.

3.5 Evaluation

To evaluate the performance of classifiers, standard 10 fold cross validation is employed. Since nature of the problem is skewed in most of the cases, accuracy may not be an effective evaluation indicator. Therefore, precision, recall and f-measure are used to evaluate the performance of classifiers. In this work, these indicators for positive class are defined as follows:

$$recall = \frac{\text{number of instances that are predicted as positive}}{\text{number of positive instances}} \quad (10)$$

$$precision = \frac{\text{number of instances that are predicted as positive}}{\text{total number of positive predictions}} \quad (11)$$

$$f\text{- measure} = \frac{2 \times precision \times recall}{precision + recall} \quad (12)$$

Precision and recall for our case is explained with following example. Suppose in a Cricket tournament, 25 matches were played overall. Suppose a classifier predicts that team *x* won (positive) 5 matches out of 8. If 3 of the

predictions were correct and remaining 2 matches that classifier predicted that x won actually lose (negative) it, than the classifier precision is 3/5 and its recall is 3/8.

4.0 EXPERIMENTS AND RESULTS

The experimentation of this work was performed using WEKA [30] tool. We have chosen WEKA due to several reasons: it has built-in state-of art feature selection, classification and evaluation methods along with text processing utilities such as tokenization, stop words removal and feature weighting (such as TF-IDF [31]). For experimentations, dataset of 109 Cricket matches is collected during the IPL2014 and CWC2015. The tweets of IPL2014 were collected using Twitter search API [20] and PHP Doom Parser [16]. For CWC2015, tweets were collected using Twitter API on daily basis during matches. The dataset is transformed into feature representation for training and evaluating the classifiers. The performance of all three classifiers is given in Table 3.

Table 3. Comparative performance of SVM, NB and LR

Classifiers	Precision	Recall	F-measure
SVM	0.893	0.880	0.877
NB	0.876	0.870	0.869
LR	0.867	0.863	0.862

The results suggest that SVM has an empirical advantage over NB and LR. It is so because SVM has many theoretical advantages over NB and LR. For example, it solves convex optimization (no local minima) problem which reliance over minimum examples and has robust internal over-fitting mechanism [41]. On the other hand, NB shows second best performance. Despite that, NB also has an advantage of being computationally efficient than others. However, it has two subtle issues; firstly, the lack of occurrence of attribute value with class label (e.g. no tweet contains predicted score for a team before the match) can cause estimated posterior probability to be zero. Secondly, since NB is a generative model, unlike SVM and LR, it models data rather than learning a decision boundary. Hence it is quite sensitive to outliers e.g. if very high score is predicted in a tweet, it shifts the mean of a class distribution and changes the probabilities of all the instances belonging to that class. Though LR has a little disadvantage in performance than other classifiers, LR has an advantage (that matters a lot in such applications) of being very efficient in big data scenarios like the one in this research. It is due to the reason that LR can easily be distributed. Another advantage of using LR is that, its output can be interpreted as probability and so the confidence of the prediction is also available, which can be useful for such applications.

Table 4. Results predicted by SVM for IPL 2014

Class	Precision	Recall	F-measure
Sun Risers Hyderabad	0.84	0.83	0.82
Mumbai Indians	0.86	0.83	0.84
Rajasthan Royals	0.90	0.88	0.901
Delhi Daredevils	0.83	0.82	0.81
Kings XI Punjab	0. 95	0.89	0.891
Kolkata Knight Riders	0.98	0.95	0.93
Royal Challengers Bangalore	0.89	0.89	0.89
Chennai Super Kings	0. 90	0.955	0.94
Average	0.89	0.88	0.87

The performance of SVM for predicting victory of individual teams in IPL2014 is shown in Table 4 where it can be seen that the proposed method can predict winner of a match with a precision of 0.89 and recall of 0.88.

The highest obtained precision was for Kolkata Knight Riders (KKR) and Kings XI Punjab (KXIP) teams, 0.98 and 0.95 respectively. IPL2014 final was played between and KKR and KXIP, so total number of tweets for these two teams were high as compared to other teams. This may be the reason that precision is high for these two teams which indicates that classifier accuracy improves with large dataset. The results of classifiers to predict the victory of top 10 teams in CWC2015 are shown in Fig. 4. It can be seen that SVM can predict victory of any of these teams with more than 0.9 precision and recall.

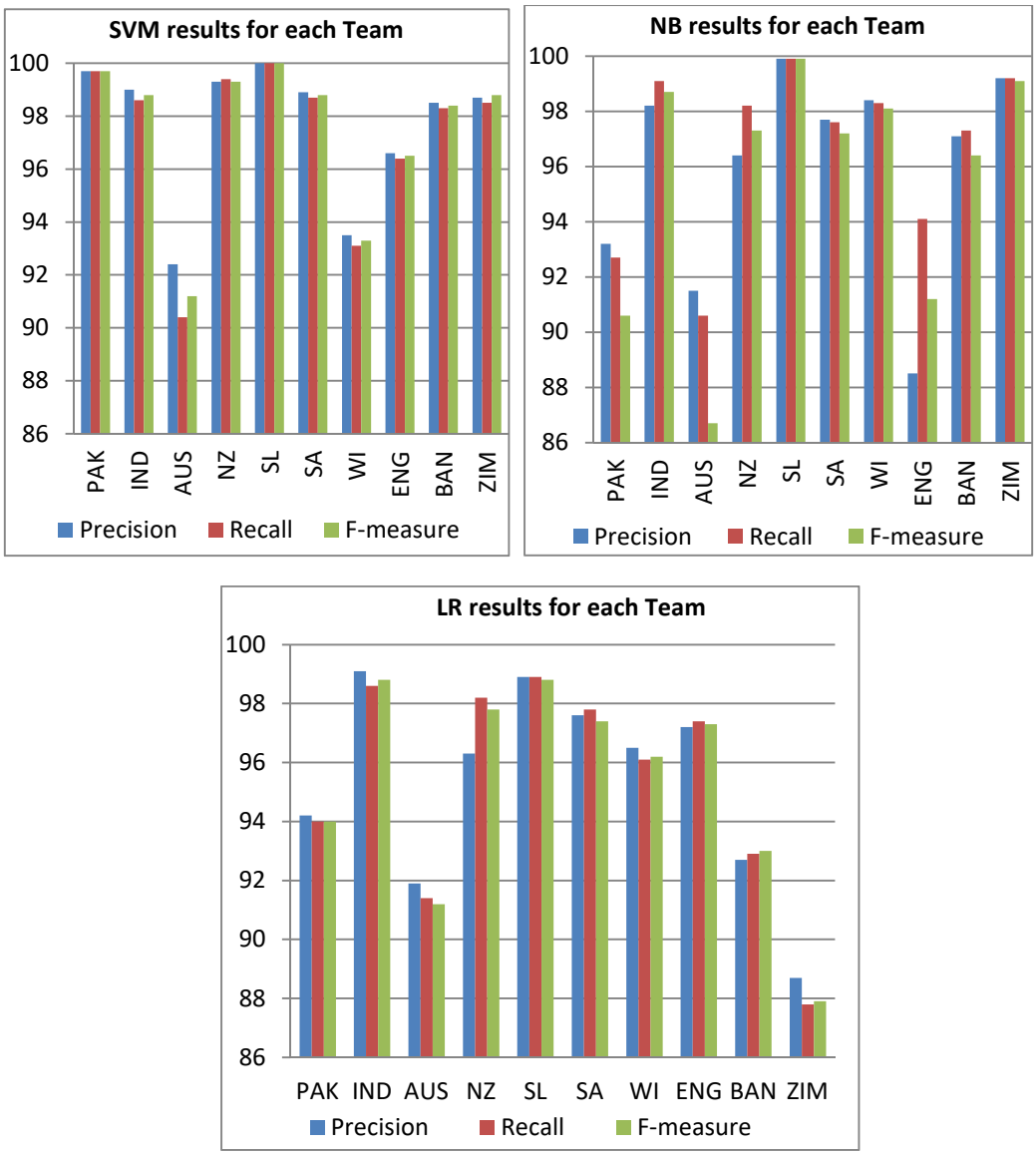

Fig. 4. Classifiers' result for each tem in CWC 2015

To better understand the effectiveness of our methodology, we also checked that whether there is any sort of profit we gained while betting on team before match. Godin *et al.* [6] achieved 60% profit betting on Soccer games in English Premier League (EPL) 2013-2014. For this, we introduced bookmaker prediction method for comparison.

Bookmaker predictions were based on the odds of 50 bookmakers from popular betting sites and we used the result with the highest probability as a prediction. The comparison was carried out in first four weeks of CWC2015 from 14 February to 13 March 2015. The accuracy of our predicted results with bookmakers predication is listed in Table 5.

Table 5. Number of correctly predicted outcomes for first 38 games in CWC 2015

Match Week	Week 1	Week 2	Week 3	Week 4	Total
No of Games	9	10	9	10	38
Bookmakers	6	7	8	9	30
Proposed Method	6	8	9	8	31

Our proposed methodology accurately predicted almost 82% of matches in first four weeks of CWC2015, 4% higher than bookmakers prediction. However, bookmakers do not reason in correct matches but in odds. Therefore, final evaluation concerns the money we theoretically could have earned (or lost) if we bet $1 on each game (from Week 1-Week 4 of CWC2015), betting a total amount of $38. With bookmakers prediction, we could have earned $23.4 and with proposed methodology we could have realized a profit of $ 25.42, which is a total profit of almost 67%. This can be explained by the fact that the proposed methodology followed a different prediction pattern than the bookmakers prediction and it was able to predict the result of two Cricket matches that had an unexpected result according to the bookmakers.

We performed paired t-test (corrected) in WEKA to check which of three classifiers are significantly better than other. Paired t-test is used to compare the result of measuring one group twice. This statistical hypothesis testing method calculates *t-value* from mean and variance of the differences between these two measures that are run several times. With *t-value* and desired significance level (0.05), the probability that these two measurements are significantly different can be obtained by looking it up from t-distribution table [25]. Comparison between classifiers for top 9 teams in CWC2015 is shown in Table 6. We selected NB as baseline classifier and each classifier was run 10 times on dataset and achieved accuracy is the mean and the standard deviation in rackets of those 10 runs.

Table 6. Classifiers comparison for 9 teams of CWC 2015.

CWC2015 Teams	NB	LR	SVM
Australia	85.85	77.40	87.55
New Zealand	81.33	85.83	93.33
India	80.24	83.64	82.14
South Africa	59.85	58.25	61.70
Pakistan	88.43	87.00	86.57
England	100	100	100
Sri Lanka	100	100	100
West Indies	80.82	79.48	79.87
Bangladesh	100	100	100

For datasets of 3 teams (Australia, New Zealand and South Africa), SVM showed better performance than NB and LR. LR showed better performance for one dataset (India) and NB showed better performance for 2 datasets (Pakistan and West Indies). The performance of the three classifiers is same for remaining three datasets. On average, SVM achieved classification accuracy of 87.90%, NB 86.28% accuracy and LR 85.73% accuracy. From the results, we can say that there exist some differences in accuracy of three classifiers but it is not statistically significant.

5.0 CONCLUSION

In this article, we checked the effectiveness of machine learning techniques when applied on collective knowledge obtained from social networks for predicting the real world events. Starting from the work done in [6, 15, 18], we used a new methodology to forecast the outcome of Cricket matches. By applying large-scale data analysis, we obtained up to 75% correct prediction. We verified our methodology on the games played in CWC2015 and IPL2014. Furthermore, we find that SVM has shown improved performance over other classifiers (NB and LR). For Cricket matches in CWC2015, our results beat the predictions of bookmakers, realizing a profit of 67%. Therefore, this methodology can be generalized for the prediction of any other Cricket matches. Our results also indicate that the social networks, such as Twitter, Facebook and blogs are as informative as professional newspaper media. Prediction systems for other games such as Hockey, Baseball, and Basketball can also be created. Such systems simply depend on the selection of right features.

REFERENCES

[1] A. Joseph, N. E. Fenton and M. Neil, "Predicting football results using Bayesian Nets and other machine learning techniques", *Knowledge-Based Systems*, Vol. 19, No. 7, pp. 544-553, 2006.

[2] T. Zia, M. P. Akhter and Q. Abbas, "Comparative study of feature selection approaches for Urdu text categorization", *Malaysian Journal of Computer Science*, Vol. 28, No. 2, pp. 93-109, 2015.

[3] A. Madani, "IPL 7 a Winner on Twitter", Available at: https://blog.twitter.com/2014/ipl-7-a-winner-on-twitter, Accessed on: 20 February 2015.

[4] A. Tumasjan, T. O. Sprenger, P.G. Sandner and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment", in *Proceedings of* the *Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178-185, 2010.

[5] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series", i*n Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp*. 122-129, 2010.

[6] F. Godin, J. Zuallaert, B. Vandersmissen, W. D. Neve and R. V. de Walle, "Beating the bookmakers: Leveraging statistics and Twitter microposts for predicting Soccer results", in *Proceedings of the KDD Workshop on Large-Scale Sports Analytics*, 2014.

[7] G. Forman, "An extensive empirical study of feature selection metrics for text classification", *The Journal of Machine Learning Research,* Vol. 3, pp. 1289-1305, 2003.

[8] H. Rue and O. Salvesen, "Prediction and retrospective analysis of Soccer matches in a league", *Journal of the Royal Statistical Society: Series D (The Statistician), Vol.* 49, No. 3, pp. 399-418, 2000.

[9] J. Bollen, H. Mao and X. J. Zeng, "Twitter mood predicts the stock market", *Journal of Computational Science*, Vol. 2, No. 1, pp. 1-8, 2011.

[10] J. Surowiecki and P. Silverman, "The wisdom of crowds", *American Journal of Physics*, Vol. 75, No. 2, pp. 190-192, 2005.

[11] K. Raj and P. Padma, "Application of association rule mining: A case study on team India", *in Proceedings of International Conference on Computer Communication and Informatics,* pp. 1-6, 2013.

[12] M. Bailey and S. R. Clarke, "Predicting the match outcome in One-Day International Cricket matches, while the game is in progress", *Journal of Sports Science and Medicine*, Vol. 5, No. 4, pp. 480-487, 2006.

[13] M. C. Wellons, "11 Predictions on the future of social media" [online], Available at: www.cnbc.com/id/102029041, Accessed on: 10 January 2015.

[14] M. I. Lali, R. U. Mustafa, K. Saleem, M. S. Nawaz, T. Zia and B. Shahzad, "Finding healthcare issues with search engine queries and social network data", *International Journal on Semantic Web and Information Systems*, Vol. 13, No. 1, pp.48-62, 2017.

[15] N. UzZaman, R. Blanco and M. Matthews, "TwitterPaul: Extracting and aggregating Twitter predictions", *arXiv preprint: 1121.6496*, 2012.

[16] PHP Simple HTML DOOM parser, Available at: http://simplehtmldom.sourceforge.net/.

[17] S. Asur and B. A. Huberman, "Predicting the future with social media", *in Proceedings of 2010 International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 492-499, 2010.

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

[19] T. B. Swartz, P. S. Gill and S. Muthukumarana, "Modeling and simulation for One-day Cricket", *Canadian Journal of Statistics,* Vol. 37, No. 2, pp. 143-160, 2009.

[20] Twitter search API, Available at: https://dev.twitter.com/rest/public/search.

[21] V. V. Sankaranarayanan, J. Sattar and L. V. Lakshmanan, "Auto-Play: A data mining approach to ODI Cricket simulation and prediction", *in Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 106*4-1072, 2014.

[22] Y. Hong and S. Skiena, "The wisdom of bookies? Sentiment analysis versus the NFL point spread", *in Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, pp. 251-254, 2010.

[23] W. L. Yeow, R. Mahmud and R. G. Raj, "An application of case-based reasoning with machine learning for forensic autopsy," *Expert Systems with Applications*, Vol. 41, No. 7, pp. 3497-3505, 2014.

[24] Raj, R.G., Abdul-Kareem, S., "Information Dissemination And Storage For Tele-Text Based Conversational Systems' Learning", *Malaysian Journal of Computer Science*, Vol. 22(2):2009. Pp. 138-159.

[25] R. R. Khorasgani, "Comparison of different classification methods", *Heart Disease*, Vol. 78, No. 81.15, pp. 83-34.

[26] E. Alwagait and B. Shahzad, "When are tweets better valued? An empirical study", *Journal of Universal Computer Science*, Vol. 20, No. 10, pp. 1511-1521.

[27] A. Qazi, H. Fayaz, A. Wadi, R. G. Raj, N.A. Rahim, W. A. Khan, "The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review", *Journal of Cleaner Production*, vol. 104, pp. 1-12, 2015. ISSN 0959-6526, http://dx.doi.org/10.1016/j.jclepro.2015.04.041.(http://www.sciencedirect.com/science/article/pii/S0959652615004096).

[28] K. Nigam, K. Mccallum S. Thrun and T. Mitchell, "Text classification from labeled and unlabeled documents using EM". *Machine Learning*, Vol.39, No. 3, pp.103-134, 2002.

[29] A. Kaluarachchi and A. Varde, "CricAI: A classification based tool to predict the outcome in ODI Cricket", In *Proceedings of 5th International Conference on Information and Automation for Sustainability, pp. 250-255, 2010*.

[30] A. Qazi, K. B. S. Syed, R. G. Raj, E. Cambria, M. Tahir, D. Alghazzawi, "A concept-level approach to the analysis of online review helpfulness", *Computers in Human Behavior*, Vol. 58, May 2016, PP. 75-81, ISSN 0747-5632, http://dx.doi.org/10.1016/j.chb.2015.12.028.

[31] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2012.

[32] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features", *in Proceedings of Tenth European Conference on Machine Learning,* pp.137-142, 1998.

[33] A. Qazi, R. G. Raj, M. Tahir, M. Waheed, S. U. R. Khan, and A. Abraham, "A Preliminary Investigation of User Perception and Behavioral Intention for Different Review Types: Customers and Designers Perspective," *The Scientific World Journal*, vol. 2014, Article ID 872929, 8 pages, 2014. doi:10.1155/2014/872929.

[34] F. C. Duckworth and A. J. Lewis, "A fair method for resetting the target in interrupted One-day Cricket matches", *The Journal of the Operational Research Society*, Vol. 49, No. 3, pp. 220–227, 1998.

[35] A. Bandulasiri, "Predicting the winner in One Day International Cricket", *Journal of Mathematical Sciences & Mathematics Education,* Vol. 3, No. 1, pp. 6–17, 2008.

[36] B. M. D. Silva and T. B Swartz, "Winning the coin toss and the home team advantage in One-Day International Cricket matches", *Department of Statistics and Operations Research, Royal Melbourne Institute of Technology*, pp. 1-15, 1998.

[37] P. E. Allsopp and S. R Clarke, "Rating teams and analyzing outcomes in One-Day and Test Cricket", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 167, No. 4, pp: 657–667, 2004.

[38] D R. Choudhury, P. Bhargava, P. Reena and S. Kain, "Use of artificial neural networks for predicting the outcome of Cricket tournaments", *International Journal of Sports Science and Engineering*, Vol. 1, No. 2, pp. 87–96, 2007.

[39] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu and B. Liu, "Predicting Flu trends using Twitter data", *in Proceedings of IEEE Conference on Computer Communications Workshops,* pp. 702-707, 2011.

[40] M. Moohebat, R.G. Raj, S.B.A.Kareem, D. Thorleuchter, "Identifying ISI-indexed articles by their lexical usage: A text analysis approach", *Journal of the Association for Information Science and Technology*, Vol. 66, No. 3, pp. 501–511. doi: 10.1002/asi.23194.

[41] T. Zia, M. S. Akram, M. S. Nawaz, B. Shahzad, A. M. Abdullatif, R.U. Mustafa and M. I. Lali, "Identification of hatred speeches on Twitter", in *Proceedings of 52nd The IRES International Conference*, pp. 27-32, 2016.