Modeling Social Media Data for Sports Analytics

Vinay Kumar Jain, Shishir Kumar
Department of Computer Science & Engineering
Jaypee University of Engineering & Guna (M.P.) India
vinay2588@gmail.com, dr.shishir@yahoo.com

ABSTRACT: *This paper presents the relationship between social media data and the outcome prediction of sport event using a dataset containing messages from Twitter. For every sport events, supporters use social media to express their opinions and performance sentiments towards players and teams. These paper uses the platform of social media for examine various meaningful inferences for a particular sports event. The outcome prediction of sports event are based on two technique, sentiment analysis and volume based approach. Important insight related to business analytics in the sports towards public interest is also examined. Proposed technique has been applied to the data collected from twitter for specific teams and games in the IPL 2015 Cricket tournament.*

Keywords: Sentiment analysis, Opinion mining, Twitter, Text classification

Received: 28 December 2016, Revised 3 February 2016, Accepted 11 February 2016

1. Introduction

Social media which provide a platform for interactive applications for creating, sharing and exchange of user contents has been a primary focus in the field of information retrieval and text mining. This platform produces massive amount of unstructured textual data with hidden user relations in real time. There are a rising number of social media tools and a rapidly growing user base across all demographics regions. The most worldwide popular social media websites, Facebook and Twitter demonstrate its explosive growth and deep influence.

According to eMarketer[1], India will account for the third-largest user base on micro-blogging site Twitter at 18.1 million by the end of this year.The research firm said growth for Twitter is heavily weighted in emerging markets, with India and Indonesia to see the most consistent growth patterns [1].

Mining the valuable attributes and contents from these social media gives us an opportunity to discover crowd wisdom. Users share their opinions towards different subjects in real time which help in analysis their personal wisdom and different viewpoints. Research which emphasizes on detecting expressions, emotions, viewpoints and private states, expressed in contents is called opining mining [2]. Opinion mining from social media is one the hottest topic in research such as in opinion polls, stock prediction, election prediction, the spread of contagious disease [2,3]. Nowadays every event of sports has it presence in social media and supporters use social media platforms to discuss issues and facts regarding team and players.

Due to rise in social media data towards sports, sports related companies are interested to mine this data and extracted meaningful inference from them which help in improving their marketing strategy for products and services. For example, many fans seeking information related to particular team or player and show their interests using by liking or commenting will becomes the potential buyer for product and services. According to a recent research from Burst Media, 45% of 18-35 year olds follow sports teams or players online, and 35% of them regularly use social media platform to comment on, tweet/retweet, share or link to online sports contents [4].

The paper contribution in the area of sports analytic using social media is as follows. Presented two techniques for outcome prediction of sport event using volume based and sentiment based technique. In concern of business analytics in the sports an insight of public interest for a popularity of a particular game event which helps companies in marketing their product and services during event. For example, team franchises paying huge amount of money to players in addition to their travel and maintenance costs during the tournament. Franchise can improve their revenues by ticket sales, advertisements, or by selling merchandise. Analytics of social media provides the popularity measures for a particular event which can be applied in variable ticket pricing. Variable pricing is popular method of charging different prices for the same seat depending upon the event. These prices are decided before the start of the tournament and they remain static throughout the tournament.

This paper presents a technique to predict event outcome from social media data using sentiment and volume based approach. The techniques have been applied towards sports event such as Indian Premier League 2015 Cricket tournament, which is one of the famous cricket tournaments in India [5]. It is an annual cricketing competition that is held in India from April-May every year between eight teams. This event is one of the widely watched events in India during 2015 and a lot of zest surrounds this contest. Brief introduction towards teams and their social media followers are given in Table 1.

IPL Teams 2015	**Official Hastags**	**Supporters**	
		Twitter[6]	Facebook[7]
Chennai Super Kings(CSK)	#CSK	1.82M	11,689,343
Mumbai Indians(MI)	#MI	1.17M	9,404,178
Royal Challengers Bangalore (RCB)	#RCB	1.16M	6,793,002
Sunrisers Hyderabad(SH)	#SRH	652 K	3,218,224
Kolkata Knight Riders(KKR)	#KKR	1.24M	12,585,989
Rajasthan Royals(RR)	#RR	687K	3,265,009
Delhi Daredevils(DD)	#DD	588K	3,098,034
Kings XI Punjab(KP)	#KP	719K	7,463,726

Table 1. Comparative analysis of supporters in Social Media

2. Related Work

There are number of methods used to forecast success (winners and losers), both for single games and team games in the sport. These events have a great uncertainty towards the results if teams and players are strong contenders. Limited research is carried out to forecast results using social media data. According to Wang [8] supporter of the team and players use social media to express their emotion and opinions .Yu et al. use twitter data for FIFA world cup 2014 to analysis the emotions users and also describe event based tweets response [9]. Some authors also studies to predict outcomes of EPL games played during the 2013-2014 season based on sentiment analysis [10, 11]. Sinha et al. used n-grams from Twitter data sets to predict outcomes of the National Football League (NFL) and compared it with other simple statistics methods [12].Lock & Nettleton, 2014 applied machine learning technique to classify tweets and also used situational variables [13]. UzZaman et al. used a framework

(TwitterPaul) to extract tweets and find the outcome of FIFA World cup 2015[14].Sentiment based technique used by Hong and Skiena [15] well predict the winner of American football (NFL).Harnessing the wisdom of the crowds from these social media data for making predictions towards sports event need more effort. Authors used Twitter for forecasting results in different domains such as in election, stock market and Box office revenues and diseases [4, 16-18].

3. Data Collection Methodology

The proposed method is started by identifying tweets and queries that are relevant for indicating the presence of IPL 2015 Cricket Tournament. For a good query for data collection, it is necessary to use relevant keywords. Relevant keywords are extracted from popular news articles and trending hastags in Twitter. This methodology gives dynamic keywords which are popular during a particular time period and helps to relate the public sentiments. Period for tweets collection is between 12 to 24 hours before the match. Examination of tweets based on proposed technique is applied for 25 matches.

Figure 1. Method of Data Collection

To retrieve tweets from Twitter, we used two APIs: Streaming API and REST API[19]. It supports short-lived connections and is rate limited. This means that there is a limit for the amount of data that can be retrieved within 24 hours. REST API access is also limited to tweets not older than a week. Alternatively, Twitter Streaming API allows high-throughput near-real-time access to various subsets of public and protected Twitter data [17].Data collection procedure is given in Fig.1.

Top hashtags during events are: #IPL, #IPL2015, #IndiaKaTyohaarBegins, #IPL8, #PepsiIPL2015, #PepsiIPL, #KKRvMI, #CSKvDD, #KXIPvRR, #CSKvRR

Every cricket match during IPL 2015 season has corresponding hastags such as #KKRvMI, #CSKvMI, #RRvKKR etc.

3.1 Data Preprocessing

Pre-processing the tweets is carried out such as stop word removal, stemming etc. We applied classification to differentiate real-time data from noise or irrelevant tweets. Thus, the purpose of this step is to decrease the amount of noise from the tweets and filter out as irrelevant tweets. Every relevant word related to IPL 2015[5] is considered as features and store in the database.

3.2 Finding Relevant Tweets

Hashtags and players names are the key features used to assign tweets to different teams. If a tweet contained players name for a particular team or the hashtags corresponding to exactly one IPL team, we assigned the tweet to that team and used it for the analysis.

Timestamps based approach is used to assign the tweets to the particular cricket match, but linking them to teams is more difficult because of presence of multiple hastags. These tweets are discarded from the data set.

Two datasets are created in which one data set is related to particular team (D_T) and second related to the corresponding match between the team (D_M).

4. Experimental Analysis

Two techniques on the dataset had been applied, firstly volume based technique in which counting of tweets containing specific keywords related to the IPL 2015 match is applied on dataset D_T , and secondly based on sentiment of users using SentiWordNet[20] sentiment engine applied on data set D_M .

Correlation of the information gathered from the twitter analysis to the match statistics is shown in Table 2.The prediction

Date	Match	Tweets Collected	Prediction		Actual Result
			Volume Based	Sentiment Based	
08-04-2015	KKR v MI	1328	KKR	MI	KKR
09-04-2015	CSK v DD	1471	CSK	CSK	CSK
10 -04- 2015	KP v RR	963	RR	RR	RR
11 -04-2015	CSK v SH	1447	CSK	CSK	CSK
11 -04- 2015	KKR v RCB	1256	KKR	KKR	RCB
12 -04- 2015	DD v RR	1125	DD	RR	RR
12 -04- 2015	MI v KP	956	MI	MI	KP
13 -04- 2015	RCB v SH	1029	RCB	RCB	RCB
14 -04- 2015	MI v RR	1152	RR	MI	RR
15 -04- 2015	KP v DD	1013	KP	KP	DD
16 -04- 2015	SH v RR	803	RR	RR	RR
17 -04- 2015	MI v CSK	1352	MI	MI	CSK
18 -04- 2015	DD v SH	985	DD	DD	DD
18 -04- 2015	KP v KKR	968	KKR	KKR	KKR
19 -04- 2015	CSK v RR	1002	CSK	CSK	RR
19 -04- 2015	RCB v MI	1365	MI	MI	MI
20 -04- 2015	DD v KKR	1125	KKR	KKR	KKR
21 -04- 2015	RR v KP	1045	RR	RR	KP
22 -04- 2015	KKR v SH	1225	KKR	KKR	SH
22 -04- 2015	CSK v RCB	1286	CSK	CSK	CSK
23 -04- 2015	MI v DD	1481	MI	MI	DD
24 -04- 2015	RCB v RR	1384	RCB	RCB	RCB
25 -04- 2015	MI v SH	1109	MI	MI	MI
25 -04- 2015	CSK v KP	950	CSK	CSK	CSK
26-04-2015	KKR v RR	1163	KKR	KKR	No result

Table 2 . Twitter analysis prediction to the actual match statistics

accuracy was used to assess the success technique in both models. This was the number of correctly predicted games divided by the number of predicted games (%) within the specified test set using the Eq. (1):

$$\text{Accuracy} = \frac{\text{no.of correctly predicted match}}{\text{total no.of predicted match}} * 100 \quad (1)$$

Volume based method have an accuracy of 60 % as compared to sentiment based technique which have 52% accuracy.Examination using McNemar test [21] is also performed to detect which method give better accuracy. McNemar test [16] which uses matched binary pairs taken from the output of two distinct predictive models:

Model 1 -Volume based and Model 2 -sentiment based. In this test each pair relates to correct and incorrect forecast. Comparison of each pairs are computed and placed into four categories:

1. a= Model 1 was correct, Model 2 was correct
2. b= Model 1 was correct, Model 2 was incorrect
3. c= Model 1 was incorrect, Model 2 was correct
4. d= Model 1 was incorrect, Model 2 was incorrect

The totals of the two discordant results (i.e. where Model 1 and Model 2 achieved different results) are input into the following McNemar formula given in Eq. (2):

$$\chi 2 = \frac{(b-c)^2}{b+c} \quad (2)$$

This X^2 value represents the McNemar [98] statistic and referred to the table of X^2 distribution [21], it shows the level of significance between two models. This 'p' value extracted from the table represents how likely it is that the difference in accuracy between two models was achieved.

McNemar chi-squared statistic is 0.333333

Corresponding p-value is 0.563703

McNemar chi-squared statistic with Yates correction of 0.5 is 0.083333

Corresponding p-value is 0.772830

McNemar chi-squared statistic with Yates correction of 1.0 is 0.000000

Corresponding p-value is 1.000000

Result using binomial exact test is 0.250000

Odds ratio equals 2.000000 with 95% Confidence interval from 0.181345 to 22.057378

In concern of business analytics in the sports some important insight of public interest for the popularity of a particular game event or players are also examined using corresponding data sets. Using count based technique applied on the data set provide two measures related to players and popularity of event is presented in Table 3.

Popularity measures extracted from social media definitely help franchise and companies in player's auctions, ticketing pricing and advertisements and hence, indirectly help companies in improving the marketing strategy.

Top Players	Popular Cricket Match
V Kohli (RCB)	#KKRvMI
RG Sharma (MI)	#KKRvCSK
AB de Villiers (RCB)	#CSKvRR
MS Dhoni (CSK)	#CSKvDD
CH Gayle (RCB)	#KXIvCSK
Harbhajan Singh (MI)	#RRvDD
AM Rahane (RR)	#KKR v SH
SL Malinga (MI)	#KXIPvRR
Yuvraj Singh (DD)	#CSK v SH
SK Raina (CSK)	#CSK v RCB

Table 3. Top player and Top Cricket Matches

5. Conclusion and Future work directions

Social media data offers unique challenges and opportunities for analysis crowd wisdom which help in prediction for sports events outcomes. This paper presented a method for analysis of tweets for the sport event IPL 2015 cricket tournament in India during March 2015 to April 2015. Period for tweets collection is between 12 to 24 hours before the match and outcome is take place before the match. Some meaningful inference is also examined in concern to sports business. Results showed volume based techniques is better perform as compared to sentiment based approach. This paper tried to contribute in the field of sport analytic using social media which have a vast amount of information hidden in it.

References

[1] eMarketer (2015) http://www.emarketer.com/Corporate/Coverage#/results/1298.

[2] Pang, B., Lee, L.(2008) Opinion mining and sentiment analysis, Foundations and Trends in *Information Retrieval* 2, 1–135.

[3] Jain,V.K., Kumar,S., (2015). An Effective App-ro-ach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter, *Procedia Computer Science*, Volume 70, p.801–807,2015.

[4] BurstMedia(2015)[http://www.postano.com/blog/how-social-media-is-changing-sports-marketing]

[5] IPL 2015[https://en.wikipedia.org/wiki/2015_Indian_Premier_League]

[6] Twitter(2015) https://twitter.com/.

[7] Facebook (2015) https://www.facebook.com/

[8] Wang, X. (2013) Applying the integrative model of behavioral prediction and attitude functions in the context of social media use while viewing mediated sports. *Computers in Human Behavior*, 29, 1538–1545.

[9] Yang Yu, Xiao Wang (2015) World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets.*Computers in Human Behavior* 48. 392–400.

[10]Godin, F., Zuallaert, J., Vandersmissen, B., Neve, W .D., Walle, R V D (2014). Beating the Bookmakers: Leveraging Statistics

and Twitter Microposts for Predicting Soccer Results, *In*: KDD Workshop *on Large-Scale Sports Analytics*, Sydney, Australia.

[11] Radosavljevic, V., Grbovic, M., Djuric, N., Bhamidipati, N (2014). Large-scale World Cup 2014 outcome prediction based on Tumblr posts, *In*: KDD Workshop *on Large-Scale Sports Analytics*, Sydney, Australia.

[12] Sinha, S., Dyer, C., Gimpel, K., Smith N (2013). Predicting the NFL using Twitter. arXiv Preprint arXiv:1310.6998, 1–11. Retrieved from http://arxiv.org/abs/1310.6998.

[13] Lock, D., Nettleton D (2014).Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports*, 10 (2) 9.

[14] UzZaman, N., Blanco, R., Matthews M (2012) .TwitterPaul: Extracting and Aggregating Twitter Predictions. *Artificial Intelligence; Physics and Society*. Retrieved from http://arxiv.org/abs/1211.6496.

[15] Hong,Y., Skiena,S.(2010), TheWisdom of Bookies? Sentiment Analysis Versus the NFL Point Spread,Proceedings of the Fourth International AAAI Conference on *Weblogs and Social Media*, Washington, D.C., May 23 – 26.

[16] Xuriguera, R (2013) Forecasting with Twitter Data. *ACM Transactions on Intelligent Systems and Technology*, 5 (1) Article 8, December 2013.

[17] Tumasjan, A., Sprenger, T.O., Sandner, G.,T., Welpe, M.I. (2011)Election forecasts with twitter how 140 characters reflect the political landscape, *Social Science Computer Review* 29, 402-418.

[18] Joshi, M., Das, D., Gimpel, K. and Smith .N.,A. (2010). Movie Reviews and Revenues: An Experiment in Text Regression, *NAACL-HLT*.

[19] Twitter (2015). https://dev.twitter.com/docs/api/1/get/search

[20] SentiWordNet (2015). http://sentiwordnet.isti.cnr.it/

[21] Dwyer, A. J (1991). Matchmaking and mcnemar in the comparison of diagnostic modalities. *Radiology*, 178 (2) 328.