Text Clustering Based Analysis of Users' Comments For Sports

Pravina Parmar[1], Prof. Trupesh Patel[2]
[1, 2] Gujarat Technological University, Ahmedabad, Gujarat, India

Abstract- *Sports Data Mining has gone viral in recent years. As started with fantasy league players and sporting enthusiasts seeking an edge in predictions, tools and techniques began to be developed to better measure both player and team performance. This type of new methods of performance measurement of a player or team is starting to get the attention of major sports franchises. Before the usage of data mining, they relied almost exclusively on human expertise. It was believed that domain experts (coaches, managers and team management) could effectively convert their collected data into usable knowledge.*

Booming of social media around the world has allowed channeling of the voice of sports fans that have essentially led to gathering and storing fan-generated, large-scale opinions about sports matches and team performance. Although research utilizing social media data for the purposes of supporting consumer market research has been increasing throughout the recent decade, there is a lack of studies using social media mining approach to improve team performance.

In this research work we are working over Fan's Demand based on Comments, particularly in Cricket in India on the various news articles related to sports. We are using Text Clustering Technique for this purpose.

Keywords- Data Mining, Text Clustering, Sports Data Mining, Social Media Mining

I. INTRODUCTION

Sports Data Mining

During the various gameplay in international level large amounts of numerical and text data is generated. In Team Sports this data is very important. Mostly these data is used by various coaches, players, team management and leagues. With the help of technology various broadcasters are also using the same data for their audience. To measure performance of any team or any player has been a handy way by using this type of data.

Video of the match and practice session has been a good source of data up to now. Video is divided into various frames and performance of particular players is extracted. However, this method of using video data is not cost efficient as data processing and analysis are complicated, computationally burdensome, and slow. With the ease of technology, quantitative approaches are applied such as using wearable devices including GPS to measure and calculating team performance based on batting, bowling and fielding etc. However, this study proposes a new approach to analyze team performance, especially focusing on the fans' perspective.

In our approach we are using fan's social media comments and expert views after a match. ''Wisdom of the Crowd'[1] has been used for Korean National football team as a case study. In their research they used Outcome-Driven Innovation (ODI) methodology, a strategy and innovation process of making product and marketing decisions, to determine potential 'opportunities' with regards to team performance to help teams reach their goals.

As any sports event provides too much data; each data represents the performance of a team or individual player. Analysis of these data is important to improve the performance by reducing the mistakes or weak points of the team. This data is also important for competitor analysis. Users Demand (Fan's Comments) is also an important source of data in sports. Analysis of the same can also help to improve the performance of a team.

II. RELATED WORK

There has been extensive work done on text mining related to sports using web news articles, webcast, or social media such as Twitter. One part of the research focuses on event detection or game analysis [1]–[4].

In a research done by , Xu et al. [2] researcher used text information provided by soccer webcasts to detect soccer events and searched for corresponding soccer event images with the time information of the detected events.

In another research performed by Jung et al. [3] proposed a method of association rule mining, event growth index, and pathfinder network analysis to extract team performance-related knowledge by using text information

provided by webcasts. In the research work for which we are extending our work is "'A Wisdom of Crowds': Social Media Mining for Soccer Match Analysis"[5].

An opportunity mining approach is proposed to identify opportunities to improve team performance based on text mining and cluster analysis. A case study of the 2018 Fédération Internationale de Football Association (FIFA) World Cup final qualification of Korea, Korea versus Uzbekistan, was conducted to explain how the proposed method works. Detailed methodology for the same work has been shown in Figure 1.0.

A well researched article "A sentiment analysis of who participates, how and why, at social media sport websites: How differently men and women write about football"[6] is performed by Marina Bagic Babac and Vedran Podobnik. This analysis uses a data collection via social media website and a sentiment analysis of the collected data. Their results show certain unexpected similarities in social media activities between male and female football fans. The detailed flow for the same is shown in figure 1.

We have also reviewed classification based technique "A unified semantic sports concepts classification as a key device for multidimensional sports analysis"[7] . In their research, researcher searches, extracts and collects all the distinct terms that are used by major basketball competitions websites, via NLP techniques to create a generic classification of concepts related to basketball actions. The generated competitions vocabularies, as well as the generic vocabulary are transformed into SKOS hierarchical classification schemes and then linked to each other through their similar terms that correspond to the same actions.

In Enhancing crowd wisdom using explainable diversity inferred from social media[6] researchers used a crowd selection approach. Tweet Classification has been done for Crowd Clustering. They have used Word2Vec for Classification. As a result they have found good results. A good F-Score of 0.93 has been achieved in Classification. There are some limitations like limited numbers of Strategy Considered (Some ignored) and the process is complex.

The next research paper we reviewed is Application of K-means Algorithm to Web Text Mining Based on Average Density Optimization [7] by FAN Guang-Ling, LIU Yu-Wei, TONG Jan-Qiang, ZHAO Sheng-Hai, NIE Zhi-Quan which was published in Journal of Digital Information Management , 2016. In their research work they have presented a method to improve web text clustering accuracy and integrity. They have applied the k-means algorithm based on average density optimization (Adk-means algorithm) to the web text clustering model. They have achieved an average 5% improvement in Accuracy found in earlier work. Their work has some limitations like No Particular Domain data has been tested alone. Higher numbers of clusters have been found in some cases in research work.

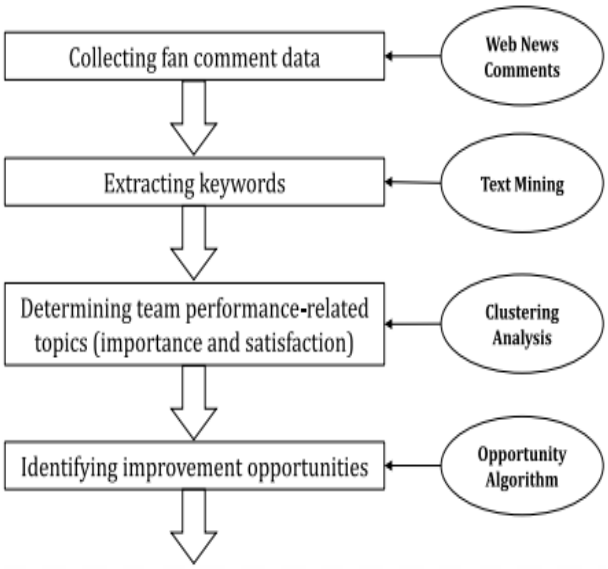

Fig. 1. Flow of Existing Research Work[1]

Following Table 1 presents the opportunities detected from the system in research[1].

Table 1. Opportunity extracted from Research [1]

No	Requirement	Opportunity Level
1	Reforming Korea Football Association	Extreme
2	Replacement of Coach	Extreme
3	Offering Coach Position to Hiddink	Solid
4	Minajae to start line up	Solid
5	Improve Player Performance	Appropriately

We also reviewed some interesting research work like "Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development"[8] to view the working and outcome of ODI. In this paper they have analyzed customer's demand and found the opportunities from them.

One other method "Analysis of Sentiments for Sports data using RapidMiner"[9] was also reviewed for another technique called RapidMiner.This research has been performed for Cricket. An overall process of feature based sentiment analysis is showcased and opinions of people are

analyzed regarding cricket matches using RapidMiner in this research.

As we are going to use K-Means + clustering method for clustering purpose, we have reviewed a research "Data Clustering with Modified K-means Algorithm" [10] by Ran Vijay Singh and M.P.S Bhatia. In their research they have used a modified K-Means algorithm.

Table 2. Related Work Reviewed

Sr No.	Title	Author & Publication, Year	Remarks
1	'A Wisdom of Crowds': Social Media Mining for Soccer Match Analysis[1]	Marina Bagic Babac, Vedran Podobnik IEEE Access , 2019	• K-Means Clustering Algorithm is used • Data used :2018 FIFA World Cup final qualification of Korea, Korea versus Uzbekistan User Comments on Social Media
2	A sentiment analysis of who participates, how and why, at social media sport websites: How differently men and women write about football [2]	Marina Bagic Babac, Vedran Podobnik Research Gate, 2018	• Mostly they have worked over emotions like Anger, Fear, Sadness, Joy from comments
3	A unified semantic sports concepts classification as a key device for multidimensional sports analysis [3]	Panagiotis-Marios Filippidis, Charalampos Dimoulas, Charalampos Bratsas, Andreas Veglis IEEE Xplore: 22 October 2018	• Collected data for Basketball using NLP. • Generated a vocabulary for Basketball Terms (Ontology) • Semantic Classification has been performed
4	Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development[4]	Junegak Joung , Kwnook Jung , Sanghyun Ko , Kwangsoo Kim Sustainability,MDPI 2019	• Outcome-Driven Innovation (ODI) method is used. • Customer Complaint has been considered as a Job. • Analyzed Complain data and detected Customer Needs using Clustering
5	Analysis of Sentiments for Sports data using RapidMiner [5]	Tanuj Pawar , Parul Kalra , Deepti Mehrotra IEEE, 2018	• Feature based sentiment analysis principle is used • Team wise Sentimental has been calculated using Rapid Miner Tool • Dataset with Cricket Comment from espn has been used
6	Enhancing crowd wisdom using explainable diversity inferred from social media[6]	Shreyansh Bhatt, Manas Gaur, Beth Bullemer, Valerie L. Shalin, Amit P. Sheth, Brandon Minnery Research Gate, 2018	• Crowd Selection Approach has been applied. • Tweet Classification has been done for Crowd Clustering • Word2Vec is used for Classification • Limited Numbers of Strategy Considered (Some ignored) • Process is Complex.
7	Application of K-means Algorithm to Web Text Mining Based on Average Density Optimization [7]	FAN Guang-Ling, LIU Yu-Wei, TONG Jun-Qiang, ZHAO Sheng-Hai, NIE Zhi-Quan Journal of Digital Information Management , 2016	• They presented a method to improve web text clustering accuracy and integrity. • k-means algorithm based on average density optimization (Adk-means algorithm) was applied to the web text clustering model • Average 5% improvement in Accuracy founded • No Particular Domain data has been tested alone. • Higher number of clusters is some cases.

After reviewing all these work we have modified our research work to get better accuracy and better results in Cricket data.

III. PROPOSED WORK

After research we have come to know with previous research done by various users that only fan's comment regarding Players or team performance may not be enough to get accurate results. Comments done on social media by fans may be biased or written with deep sentiment that may not be applicable in the current situation of match. We also added expert views given after the Match in various News Paper or on social media. We have combined the dataset of Fan's Comment and Expert Views. Figure 2.0 explains the steps for our proposed system.

Pseudo code :

Step 1: Collect User Comments from News related to Sports
Step 2: Clean the Collected Data
Step 3: Extract Keywords using TF/IDF
Step 4: Remove irrelevant keywords
Step 5: Generate Dataset
Step 6: Apply Modified K-Means Algorithm for Clustering
Step 7: Identify Fan's Demand from Each Cluster
Step 8: Based on Frequency extract important Clusters
Step 9: Using delphi method calculate satisfaction level
Step 10: Apply Opportunity Algorithm

We have divided our work in 7 Phases that are explained like Phase1 contains Data collection, which is done by API and Crawling methods. Phase 2 is cleaning process and keyword extraction. Phase 3 includes Structured Dataset Generation from gathered data collection. In Phase 4 we have applied the Modified Clustering method to make better clusters from the dataset. Phase 5 we have extracted Users Demand from each cluster. In the next Phase we have calculated Satisfaction ratio. And at the last phase Opportunity algorithm is applied.

In the first phase we have collected data using Tweepy API and News RSS. We have also used scraping for extracting comments from the news. In the second phase we have applied text cleaning methods. In this phase we have used NLP for performing TF/IDF. After preprocessing data we have generated structured dataset using collected data. We have applied the K-Means algorithm with modification in distance calculation. In the next phase fan's demand is extracted from generated clusters. To measure satisfaction level we have applied the Delphi method and last Opportunity algorithm is applied.

We have implemented our proposed system using Python 3.0. We have applied various libraries of python to perform the data extraction and algorithm processing.

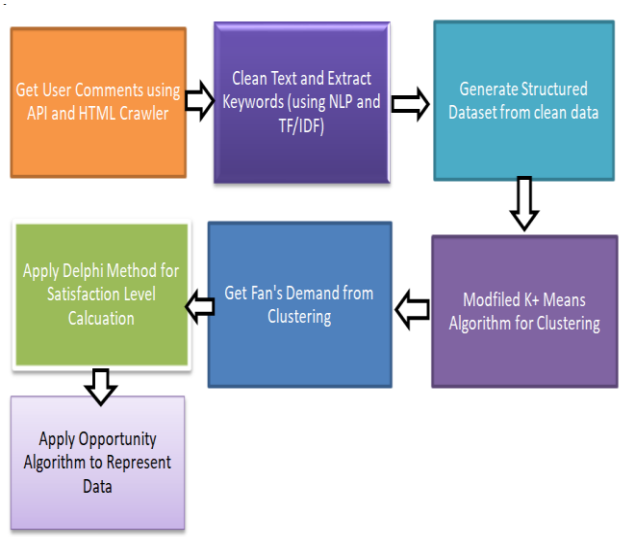

Fig. 2. Proposed Model for Our System

IV. RESULT

Experimental Setup

For our research work implementation we have used Python 3.0. We have set up our proposed research work on Windows 10 Operating System with 16 GB of Internal Memory. For collecting data from twitter we have applied Tweepy API. For various data manipulation other libraries like Panda, Article , makecluster etc applied.

As we have collected tweets from ODI Cricket World Cup 2019 which was held from May 2019 to July 2019. We have collected tweets based on the hashtag CWC19 and IndianCricket.We have collected nearly 6043 tweets regarding these hashtags and 150 news articles related to Cricket for our research purpose.

Table 3. Frequent Terms in Tweets

Sr No.	Frequents Words	Tweets Count
1	Kohli	990
2	IndianCricket	3879
3	Virat	674
4	Dhoni	1335
5	Rohit	518
6	Team India	774

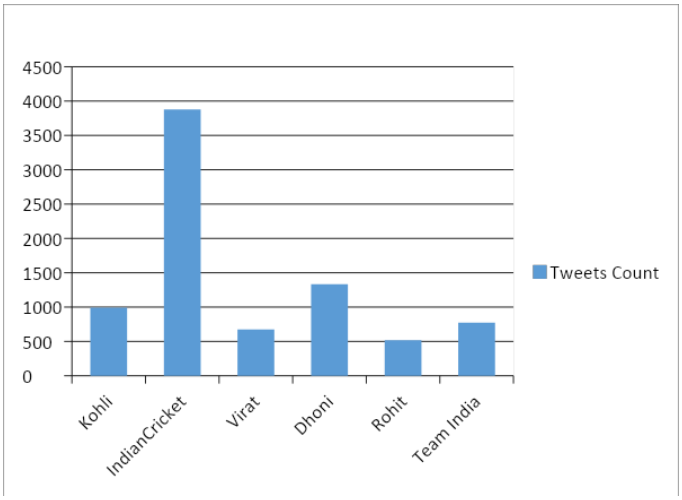

Fig. 3. Graph for Tweet Count

We have generated clusters based on text data and founded top clusters regarding topic :

Table 4. Opportunity Found from Fan's Comment

Poor Umpiring
New Zealand Deserves
Controversial Decision
Proud of Rohit
Well Played Team India

V. CONCLUSION

As we have collected and analyzed Indian Cricket fans' demands from social media; we have summarized that Most of Indian Cricket fans are satisfied with team performance and most of them have no complaint about performance. But they are very disappointing regarding decisions taken by umpires. They also believe that New Zealand was deserving the title. In future this work can be extended to find out performance of individual sportsmen or for particular match team performance. Other clustering techniques can be applied for the same. As earlier work was applying only social media data; which was not sufficient to extract all the demands from users particularly in India. Data collected from various news and Expert Columns helped our proposed work to be better than earlier systems.

REFERENCES

[1] Y. Kim and M. Kim, "'A Wisdom of Crowds': Social Media Mining for Soccer Match Analysis," in *IEEE Access*, vol. 7, pp. 52634-52639, 2019. doi: 10.1109/ACCESS.2019.2912009

[2] Bagic Babac, Marina & Podobnik, Vedran. (2016). A sentiment analysis of who participates, how and why, at social media sport websites: How differently men and

women write about football. Online Information Review. 40. 814-833. 10.1108/OIR-02-2016-0050.

[3] P. Filippidis, C. Dimoulas, C. Bratsas and A. Veglis, "A unified semantic sports concepts classification as a key device for multidimensional sports analysis," *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Zaragoza, 2018, pp. 107-112.

[4] Junegak Joung,Kiwook Jung,Sanghyun Ko, Kwangsoo Kim , "Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development ", MDPI 2019

[5] T. Pawar, P. Kalra and D. Mehrotra, "Analysis of Sentiments for Sports data using RapidMiner," *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, Bangalore, India, 2018, pp. 625-628. doi: 10.1109/ICGCIoT.2018.8752989

[6] Bhatt, Shreyansh & Gaur, Manas & Bullemer, Beth & Shalin, Valerie & Sheth, Amit & Minnery, Brandon. (2018). Enhancing Crowd Wisdom Using Explainable Diversity Inferred from Social Media. 293-300. 10.1109/WI.2018.00-77.

[7] FAN Guang-Ling, LIU Yu-Wei, TONG Jan-Qiang, ZHAO Sheng-Hai, NIE Zhi-Quan," Application of K-means Algorithm to Web Text Mining Based on Average Density Optimization" Journal of Digital Information Management, 2016

[8] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, ''Live sports event detection based on broadcast video and web-casting text,'' in Proc. 14th ACM Int. Conf. Multimedia, 2006, pp. 221–230.

[9] H.-S. Jung, J.-U. Lee, J.-H. Yu, H.-S. Lee, and D.-H. Park, ''In-depth analysis of soccer game via webcast and text mining,'' J. Korea Contents Assoc., vol. 11, no. 10, pp. 59–68, Nov. 2011. [Online]. Available: https://www.dbpia.co.kr/Journal/ArticleDetail/NODE01713822

[10] J. Nichols, J. Mahmud, and C. Drews, ''Summarizing sporting events using twitter,'' in Proc. ACM Int. Conf. Intell. User Interfaces, 2012, pp. 189–198.

[11] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, ''Using webcast text for semantic event detection in broadcast sports video,'' IEEE Trans. Multimedia, vol. 10, no. 7, pp. 1342–1355, Nov. 2008. [Online]. Available: http://ieeexplore.ieee.org/document/4668533/

[12] Bagic Babac, Marina & Podobnik, Vedran. (2016). A sentiment analysis of who participates, how and why, at social media sport websites: How differently men and women write about football. Online Information Review. 40. 814-833. 10.1108/OIR-02-2016-0050.

[13] S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan, Human as RealTime Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games. Houston, TX, USA: Rice Univ., 2011. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1106/1106.4300.pdf