Sports Match Analysis using Social Media Mining

A PROJECT REPORT

submitted by

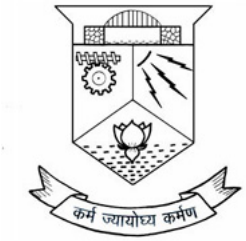
ALFIN WILLIAM
LTVE17MCA058

to

the APJ Abdul Kalam Technological University
in partial fullfilment of the requirements for the award of the degree

of

Master of Computer Applications



Department of Computer Applications
College of Engineering Trivandrum
Trivandrum-695016

APRIL 2020

DECLARATION

I undersigned hereby declare that the project report Sports Match Analysis using Social Media Mining, submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Sri.Jose T Joseph, Assoc.Professor. This submission represents my ideas in my words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity as directed in the ethics policy of the college and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title.

Place : Trivandrum
Date : Alfin William

DEPARTMENT OF COMPUTER APPLICATIONS

COLLEGE OF ENGINEERING TRIVANDRUM

CERTIFICATE

This is to certify that the report entitled **Sports Match Analysis using Social Media Mining** submitted by **Alfin William** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Computer Applications is a bonafide record of the project work carried out by him under my guidance and supervision. This report in any form has not been submitted to any University or Institute for any purpose.

Head of the Dept

Project Guide

ACKNOWLEDGEMENT

If words are considered as symbols of approval and tokens of acknowledgement, then let words play the heralding role of expressing my gratitude.

First of all, I would like to thank God almighty for bestowing me with wisdom,courage and perseverance which had helped me to complete this project ***Sports Match Analysis using Social Media Mining***. This project has been a reality as a result of the help given by a large number of well wishers.

I am extremely thankful to **Dr Jiji C V**, Principal, College of Engineering Trivandrum for providing me with the best facilities and atmosphere which was necessary for the successful completion of this project.

I would like to remember with gratitude **Dr. Sabitha S**, Head Of Department Department of Computer Applications, College of Engineering, Trivandrum for the encouragement and guidance rendered.

I express my sincere thanks to **Sri.Jose T Joseph,** Assoc.Professor, Department of Computer Applications, College of Engineering Trivandrum for his valuable guidance, support and advice that aided in the successful completion of my project.

Finally, I wish to express my sincere gratitude to all my friends, who directly or indirectly contributed in this venture.

Alfin William

ABSTRACT

The increase of social media utilisation has allowed streaming the voice of sports fans that have essentially lead to storing fan-generated, large-scale opinions about sports match and team performance. Although research utilizing social media data for the consumer market studies have sharply risen in the recent decade, there is a lack of studies using social media mining approach to improve team performance. In this project, an opportunity mining approach is fullfilled to identify opportunities to improve team performance based on text mining and cluster analysis. This mining approach utilises proven data mining techniques like KNN and TF-IDF to effectively extract and categorize fan suggestions and use it to enhance overall team performance.

One of the most popular and state-of-art methods for team performance analysis is to use video data that are captured from on-field cameras and crunched into thousands of data points per second by providing each players' performance metrics, such as player speed, position and possession time. However, this method of using video data is not cost efficient as data processing and analysis are complicated, computationally burdensome, and slow. More recently, quantitative approaches are applied such as using wearable device including GPS to measuring and calculating team performance based on running speed, distance, time, etc.

However, this project analyses a new approach to analyze team performance, especially focusing on the fans' perspective. It focuses on the value of the fans as a strategic partner in addressing some of challenges that a team has. In the fields of business and management, the expertise, ingenuity, and creativity of individual members of the public are harnessed as an innovative problem solving approach. In this regards, this study suggests finding the factors associated with winning or losing from outside of the field.

Contents

List of Figures

Chapter 1

Introduction

Aside from great talent, exceptional teamwork, and dedicated training, numerous sports athletes, coaches, managers, teams, and leagues take advantage of Big Data as it has potential to provide insights regarding the critical factors associated with winning or losing.

One of the most popular and state-of-art methods for team performance analysis is to use video data that are captured from on-field cameras and crunched into thousands of data points per second by providing each players' performance metrics, such as player speed, position and possession time.

However, this method of using video data is not cost efficient as data processing and analysis are complicated, computationally burdensome, and slow. More recently, quantitative approaches are applied such as using wearable device including GPS to measuring and calculating team performance based on running speed, distance, time, etc. However, this project works on a new approach to analyze team performance, especially focusing on the fans' perspective. It focuses on the value of the fans as a strategic partner in addressing some of challenges that a team has. Before, during and after a sporting event, fans share their opinions and unique analysis of the match. Moreover, evaluation and analysis by the experts and sports analysts provide not only useful information and knowledge of team performances but also wisdom to win.

Thus, the comments, threads, and even tweets written by the public can be considered as a unique source of data to assess team performance which is essentially equivalent to "Wisdom of the Crowd".

Chapter 2

Literature Review

AUTHOR	PAPER TITLE	YEAR OF PUBLISHING	METHODOLOGY
C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan	Live sports event detection on broadcast video and web-casting text	2006	Using 4 modules: live text/video capturing, live text analysis, live video analysis, and live text/video alignment.
Young-seok kim and Mijung kim	'A Wisdom of Crowds': Social Media Mining for Soccer Match Analysis	2019	Web scraping & API of collection, Clustering algorithms for organising and retrieval of data
Nichols, J. Mahmud, and C. Drews	Summarizing sporting events using twitter	2012	Unsupervised algorithm for generating a textual summary of events & comparison with human results.

Chapter 3

Problem Definition

One of the conventional and active methods for team performance analysis is to use real time video data that are captured from on-field stadium cameras and crunched into thousands of data points per second which provides each players' performance metrics, such as player speed, position and possession time. Video data can be analysed to pinpoint the performance figures of individual players and how their game statistics improved the overall outcome of the game. This is an data extensive technique that indeed provide rich content for future analysis and team improvement.

However, this method of using video data is not cost efficient as data processing and analysis are complicated, computationally burdensome, and slow. More recently, IoT devices are incorporated are applied such as using wearable device including GPS in smart watches to measuring and calculating team performance based on running speed, distance, time of a player in real time. Wearable devices needs favourable conditions , needs constant monitoring , maintenance and has constant risk of failure. This constant economical burden of IoT devices are a major step down as if their fail during a critical event of the match , the entire data collected during that time frame is lost forever and is not achievable by any other means.

This current situation should be overcame by introducing an error-free data collection mechanism that can be depended upon on crucial events and which comes with a minimal risk of failure.

Chapter 4

Motivation

Previously, team performance analysis was computed using video data that were captured from on-field cameras and crunched into thousands of data points per second by providing each players' performance metrics, such as player speed, position and possession time. This method indeed has its own flaws including larger economical cost , higher feasibility issues and related complexity.

Another method was quantitative approaches such as using wearable device including GPS and multiple sensors including accelerometers , compass and gyroscope to measuring and calculating realtime individual player performance based on their running speed, distance, time cadence.This method included hardware components that have a realtively higher chance of failure during critical events that may lead to loss of valuable data that were to be used for future computations to calculate team performance.

This motivated for a practical and feasible approach that works by capturing large real time internet crowd sourced data related to the sports event and computing its various properties and extracting useful information from it.

Chapter 5

Objectives and Contribution

The main objective of this project is to use crowd sourced data from various social media related to a specific sports event that is procured using web APIs from official social media vendors like Reddit , Facebook and Twitter and then process this data in such a way that the relevant information regarding team performance and factors regarding the overall form of the sports team can be evaluated from the fan's perspective and use that data to improve future team matches.

Objectives are categorised into three phases , namely , Data Collection , Data processing and Final data representation phase. In this data collection phase, relevant data related to a specific sports event is collected and organised for data processing phase. In data processing phase , the objective is to compute the collected data and extract relevant information like the overall satisfaction of fans relative to the specific sport event and various individual key factors in that game the fans primarily talked about and how it affected the overall game performance. The objective of final data representation phase is to effectively plot this results to end user.

The contribution to this area of research is intended to be to reveal the benefits and elevate the use of internet user content for sports match or any other specific event related processing. Analysing an event from the eyes of its audience and processing that data can reveal hidden details and points that can be used to improve the efficiency of future events of that domain.

Chapter 6

Requirement Analysis

6.1 Purpose

Millions of tweets , threads and comments are generated real time during a live sports match.This large pool of effective information is left untouched to be forgotten about.The purpose is to retrieve and analyse this unique crowd provided data , analyse patterns in them and effectively use that information as strategies in successive matches to improve overall team performance.

6.2 Overall Description

Around the world , hundreds of sports matches of various categories occurs in a day. This matches are well documented using visual and audio means and are analysed accordingly. But an efficent alternative to this conventional economically inefficient methods are the need of the hour.

This methods of using visual and audio analysis of data are proven to be less cost efficient as data processing and analysis are complicated, computationally burdensome, and slow.

Another point is that other more recent methods like using wearable devices to measure player statistics real time needs favourable conditions , needs constant monitoring , maintenance and comes with a constant risk of failure.

Last but not the least,conventional data collection methods heavily depends upon manual human intervention and a chance of missing a significant important event is statistically high as human error can occur in any time of the event.

Sports Match Analysis using Social Media Mining is thus thus need of the century.Using instant web scraping methodologies, we can collect million of fan comments and conclusions real time from social media and compute and sort that data corpus to extract future strategies and suggestions that can significantly improve team performance in successive matches.

The major objectives behind this work are as follows:

- Instant data collection of fan comments that analyse and critic the sports match real time.
- Connecting with fan emotions and accepting feedback from them.
- Using strategies and suggestions given by fans to improve team performance for future matches.

6.3 Overall Description

6.3.1 Hardware Requirements

- Intel Core i3 or equivalent processor
- 4 GB or more RAM
- 750 MHZ or more CPU Speed
- 500 GB or more hard disk space
- Stable Internet Connection

6.3.2 Software Requirements

- Linux/Windows

6.4 Functional Requirements

Functional requirements outline the intended behaviour of the system. This behaviour may be denoted as tasks or functions that the specified system is intended to perform. The proposed system consists of the following parts. They are given below:

6.4.1 Desktop Application Interface

A Desktop Application Interface facilitates the interaction of the users with the system. The interface is quite simple.It can help users submit the thread or tweets link to commence data collection and computation.

6.5 Non Functional Requirements

Non-Functional requirements define the general qualities of the software product. Non-functional requirement is in effect a constraint placed on the system

or the development process. They are usually associated with product descriptions such as maintainability, usability, portability, etc. It mainly limits the solutions for the problem. The solution should be good enough to meet the non-functional requirements.

6.6 Performance Requirements

- Accuracy: Accuracy in the functioning and the nature of user-friendliness should be maintained in the system.
- Speed: The system must have speed at which content is delivered to users, and how responsive the system is.

6.7 Quality Requirements

- Transparency: The system provides correct data to all participants
- Scalability: The software will meet all of the functional requirements.
- Maintainability: The system should be maintainable. It should keep backups to atone for system failures and should log its activities periodically.
- Reliability: The acceptable threshold for the downtime should be as long as possible. i.e.mean time between failures should be as large as possible. And if the system is broken, the time required to get the system back up again should be minimum. .

Chapter 7

Methodology And System Design

7.1 Overall Design

Our system is a data mining application that retrieves crowd sourced social media data and extracts useful information and conclusions from it.

7.2 User Interfaces

One of the main aims while designing the system was to abstract as much lower level details of the system as possible from the user. This system provides a web interface for its users. The interface is developed using PyQt framework.

7.3 System Design

The only technology on earth today that could handle all these problems and provide us with immutable, verifiable and trustworthy certificates is 'Data Mining'. The proposed system uses the public data mining methodologies called TF IDF algorithms and K means clustering to mine information out of primitive data. Here the focus is on solving lack of information regarding the constant pace and performance measure of a team and how it can be improved.

7.4 Data Collection phase

Web forum comments related to a specific sports match should be collected as the primary step of the proposed approach. The collected material should consist of online comment data generated by sports fans on web forums; our approach identifies major team performance-related topics and criticisms that are currently being discussed by fans and the satisfaction analysis of the team's performance in overall match is performed from a fan-centered perspective therefore comments from fans should be the primary source of data. Collection of online comments may be conducted with techniques used to collect large-scale social media data such as web crawling via interfacing with application programming interfaces (APIs).

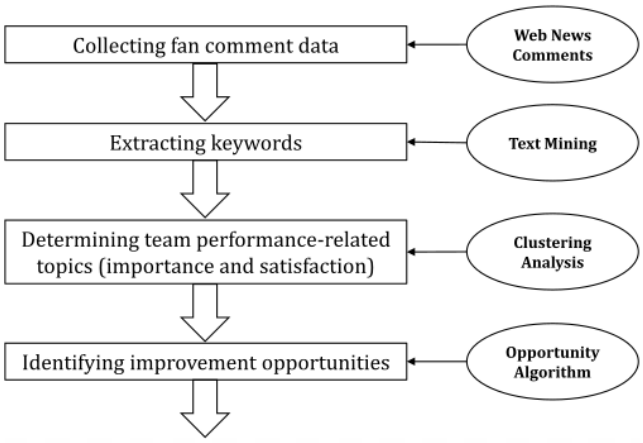


Figure 7.1: Multiple Stages of User Comments Retrieval and Processing

7.4.1 Social Media API

Social media APIs of various vendors including official Reddit and twitter API is used to collect specific user comments and tweets from a thread using a hashtag like keyword or unique thread id in case of Reddit. Twitter is an American microblogging and social network- ing service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read them. Users access Twitter through its website interface, through Short Message Service (SMS) or its mobile-device application software. There has been a steady increase in the sheer volume of user tweets proportional to the increase in years. This is largely untouched pool of data that contains insights to many unrecognised patterns and information.

7.4.2 TF - IDF Algorithm

After the extraction/collection of a set of fan comments of the target sports match, the next step of our approach is to extract keywords (or key phrases) from the fan comments to structure each of the comments. This can be performed using Term Frequency - Inverse Document Frequency Algorithm. Term frequency – Inverse Document Frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today; 83 percent of text-based recommender systems in digital libraries use tf–idf.

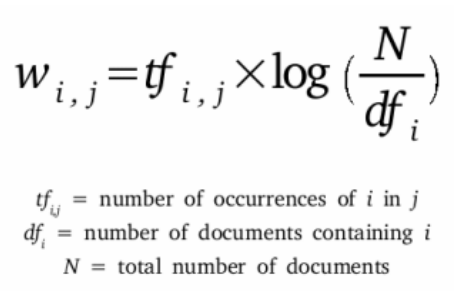


Figure 7.2: TF - IDF algorithm

7.4.3 Tokenization of data

Tokenization is a very common task in NLP, it is basically a task of chopping a character into pieces, called as token, and throwing away the certain characters at the same time, like punctuation. Tokenization includes word and sentence tokenization . Tokenization a data corpus allows easy computation and manipulation of its relevant data and ignoring the less needed semantic and language jargons.Tokenization of a word corpus is seen as an initial step in performing text classification, intelligent chatbot, sentimental analysis, language translation, etc. It is vital to understand the pattern in the text to achieve the above-stated purpose. These tokens are very useful for finding such patterns as well as is considered as a base step for processes like stemming and lemmatization.

7.4.4 Removing Stop Words

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. This collection of words are grammatically necessary in a sentence but is less useful during processes like data mining. The initial stage of preprocessing of data include removing this stop words from the data corpus so as to simplify the dataset and perform efficient computational tasks in it. Removal of stop words are generally performed using python’s Natural Language Toolkit Library.

```
> stopwords("english")
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
```

Figure 7.3: Stopwords in English Language

7.5 Data Processing Phase

After collecting and preprocessing the data, the next step in the approach is to define team performance-related topics and compute the importance value of each topic using clustering algorithm. Clustering algorithms such as hierarchical clustering or k-means are applied to a similarity matrix.

7.5.1 Hierarchical Clustering

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: Agglomerative, also known as bottom-up approach or hi- erarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data. Divisive , also known as top-down approach. This algorithm also does not require to prespecify the number of clusters.

7.5.2 K-Means Clustering

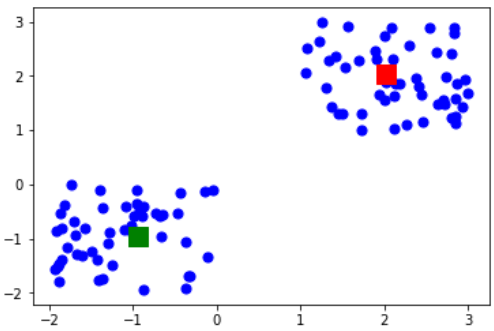


Figure 7.4: K-means clustering with cluster centroids

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serv- ing as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-Means minimizes within- cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes

squared errors, whereas only the geometric median minimizes Euclidean distances. Better Euclidean solutions can for example be found using k-medians and k-medoids.

7.5.3 Sentiment Analysis

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

Sentimental Analysis can effectively analyse the fan com- ments emotions and hence derive their satisfaction and ap- proval regarding individual player performances and about the overall Sports match in general.There are mainly two approaches for performing sentiment analysis. Lexicon-based , count number of positive and negative words in given text and the larger count will be the sentiment of text. Machine learning based approach , Develop a classification model, which is trained using the pre-labeled dataset of positive, negative, and neutral.

Multi-Nomial Naive Bayes Classification using scikit-learn is the optimal method to classify fan comments in this specific scenario.

7.6 Result Representation Phase

To effectively represent the derived findings , an optimal and lightweight python GUI library should be used that can accommodate the entire results of most relevant fan comments and their sentiment analysis and cluster classifications without causing computational and resource overhead.The most optimal GUI framework for achieving this objective is ipywidgets.

7.6.1 Ipywidgets

Ipywidgets are interactive HTML widgets for Jupyter notebooks, JupyterLab and the IPython kernel.

Notebooks come alive when interactive widgets are used. Users gain control of their data and can visualize changes in the data.Ipywidgets help achieve that simplicity to otherwise complex functions and streamlines data flow.

Learning becomes an immersive, fun experience when using ipywidgets. Researchers can easily see how changing inputs to a model impact the results and make it more straightforward for normal users.

7.7 Data Flow Diagrams

These diagrams gives a clear picture about the privileges of each user. Also the entire working flow was specified in this. The DFDs are as follows:

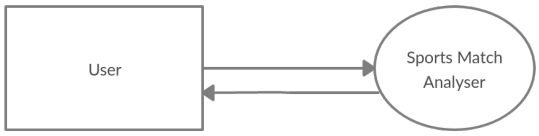


Figure 7.5: Level 0 Data Flow

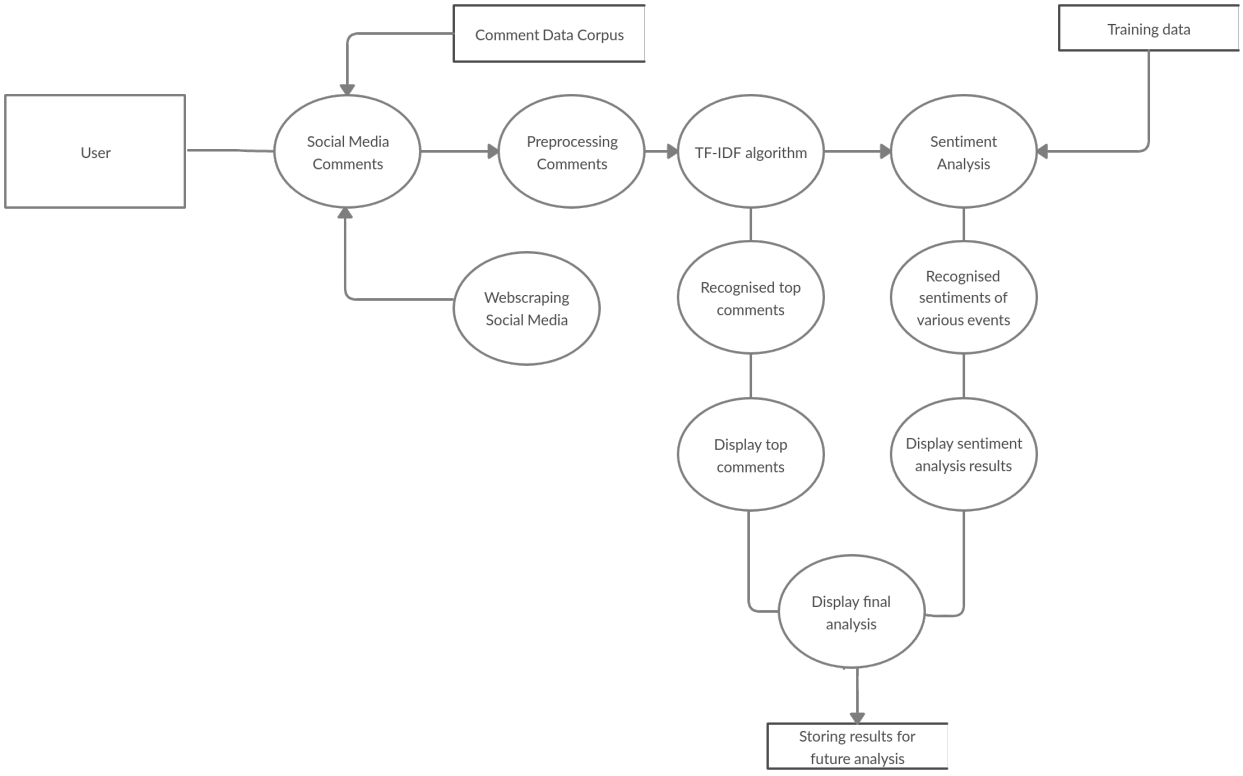


Figure 7.6: Level 1 Data Flow

Chapter 8

Results And Discussion

8.1 Social Media API Data Results

The project was developed using python as the programming language , Juypter Notebook as IDE and various other tools and packages for code formatting and compilation. Collection of social media comments involved using Social media API's of official vendors like Twitter , Reddit and Facebook. The specific thread link in reddit or hashtag in twitter was used as an identifier for specifying user comments and python modules like Tweepy and Praw was utilised for data collection.

It was noted that the comment quality of reddit was superior to twitter tweets containing more factual information as suggestions related to team performance. Thus , more social media comments from reddit proportional to twitter was analysed in data collection phase.

```
                                                comment
0     So this is the worst series of Kohli like in y...
1                              Saini ODI average at 53 😂
2     SANTNER OVERRATED\n\nNEESHAM OUTDATED\n\nLONG ...
3     Kohli running with Shaw is going to be interes...
4                                             [deleted]
...                                                 ...
4640                    OK, thanks for your feedback.
4641  Fair enough. I just still try to remember they...
4642                                Each to his own man
4643                                                 😂😂
4644                                               Cool

[4645 rows x 1 columns]
```

Figure 8.1: Reddit Data Extraction using PRAW

The collected data corpus underwent semantic preprocessing like word tokenization and sentence tokenization to make the corpus more accurate for future data mining computations. Stopword Removal was also performed on the final data corpus.

8.2 Data Mining and Semantic Analysis Results

In data processing phase , Term Frequency - Inverse Document Frequency algorithm was applied on the data corpus to find the most commonly talked about

terms by the fans related to that specific sports event. It was found out that fans mostly talked about a specific sports men or about an individual event in the game in general. The result of this computation was stored for future retrieval.

```
bhai_saheb  -  0.04099023818138398
15bfeqj1kl  -  0.02049511909069199
deba  -  0.02049511909069199
debates  -  0.02049511909069199
opened  -  0.04099023818138398
imvkohli  -  0.06148535727207597
sledge  -  0.04099023818138398
aussies  -  0.04099023818138398
matthewwade  -  0.04099023818138398
wicketkeeper  -  0.029164892066626372
australian  -  0.04099023818138398
words  -  0.04099023818138398
recent  -  0.04099023818138398
sportsflashes  -  0.02049511909069199
fn0uegxgss  -  0.02049511909069199
finale  -  0.02049511909069199
cricket  -  0.07291223016656594
h33qfkeafq  -  0.02049511909069199
```

Figure 8.2: TF - IDF results with weight scores

After calculating the TF - IDF weights , they were compared with the keywords of the original web scraped fan comments and the resultant top keywords and comments related to it was identified and displayed. This top keywords the events / instances / sports persons the fans acknowledged more in the respective match. Also the detailed comments containing this keywords can also be viewed and analysed to arrive at a conclusion how the importance factors that led to the match failure / success.

```
Fans mostly talked about:
```

series	india	kohli	saini	like	bowling	rahul	cricket	bumrah

Figure 8.3: Top comments calculated using TF - IDF Weights

```
Top comments about : bowling

 Hit a century belting the piss out of a near full strength English bowling unit in a warm up.
 All 3 of them capable of building innings as well as hitting the big shots when needed"
 Things going so good for KL Rahul, he should open the bowling and see if he can get a wicket or two.
 "Only bowling is keeping grandmom in the side.
 Were finally not opening the bowling with Bennett, good choice
 I see it's India's time to attempt suicidal runs
 "So glad to see no 4&5 batsmen actually scoring runs
```

Figure 8.4: Highlighted comments of a keyword bowling

K - Means Clustering algorithm was applied on the web scraped data-set to classify it to sub domains of the sports match , prominently into Batting , Bowling Fielding. This clusters were then matched with the TF - IDF words to find the most talked about keywords by fans in each cluster. This was sorted and stored for future data representation.

Figure 8.5: Top comments of Bowling Cluster

On the basis of results , it was established that the top comments of each cluster was mostly similar to each other due to the similarity in their domain. The results would have been much distinguishable if data of different domains were to be evaluated.

Sentiment Analysis was also performed on the comment corpus to find the fan satisfaction in the current sports event and individual top ranked user comments also underwent sentiment analysis to analyse how fans are emotionally reacting to that particular event.

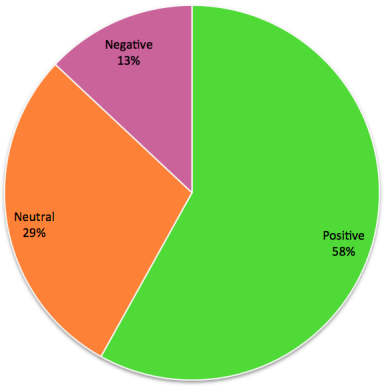


Figure 8.6: Sentiment Analysis Result of total fan comments

Domain-based sentiment analysis was also performed on the previously stored data clusters how find out level of satisfaction of fans in the three main clusters , Batting , Bowling Fielding. Like retrieval of top comments , sentimental analysis measurement of various clusters also resulted them having similar values due their nature of sharing similar domain ie Cricket in general. This results were also stored for future data representation.

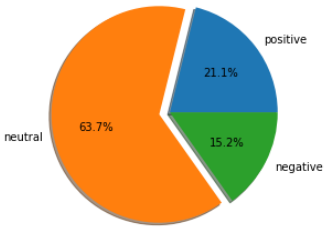


Figure 8.7: Batting Cluster

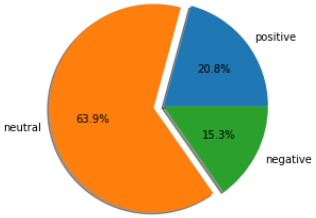


Figure 8.8: Bowling Cluster

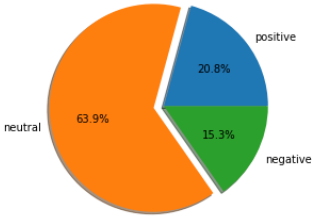


Figure 8.9: Fielding Cluster

All the collected information ie the top most relevant comments and sentiment analysis resulted was stored to a secondary storage for future analysis and retrieval. The cluster data sets and primary web scraped data results are always stored as CSV formatted files for future analysis and retrieval.

Chapter 9

Conclusion and Future Work

This project completion realises the untouched scope of utilization of online fan comments for the identification of sports team performance improvement opportunities through combining text mining and opportunity algorithm as the core methodology rather than using conventional methods like live video or IoT peripheral devices. Regarding the specific steps of the approach, each team performance-related topics from fans' perspective was defined by clustering of the online fan comment data. Following the clustering, each topic's measure of importance and satisfaction were computed. Specifically, the concept underlying the computation of importance (of the topic) is contribution (i.e., number of times mentioned) and the concept underling the computation of satisfaction (of the topic) is sentiment (i.e., positive vs. negative). As the final step, the results were organised and shown to the end - user using a GUI application interface.

Future scope includes expanding this functionality to other social media's using their official API's. Also , automation and streamlining of the entire process using continous integration / continous deployement pipeline can be considered. Realtime collection and computation of fan comments during the event can be realised using Cloud computing and historic and large data corpuses of fan comments can be stored and computed upon using Big Data methodologies.

References

1. **J**. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," in Proc. ACM Int. Conf. Intell. User Interfaces, 2012, pp. 189–198.

2. **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

3. **T**.-M. Yen, Y.-C. Chung, and C.-H. Tsai, "Business opportunity algo- rithm for ISO 9001: 2000 customer satisfaction management structure," Jan. 2007.

4. **S**. G. Park, G. S. Won, and S. W. Lee, "Web news comment-based sentiment analysis of the South Korean national team members in the 2014 Brazil World Cup," Korean J. Sport Manage., vol. 20, vol. 2, pp. 13–28, Apr. 2015.

5. **C**.-M. Chen and L.-H. Chen, "A novel approach for semantic event extraction from sports webcast text," Multimedia Tools Appl., vol. 71, no. 3, pp. 1937–1952, Aug. 2014

6. **M**. S. T. Deokar, "Text documents clustering using K means algorithm," Int. J. Technol. Eng. Sci., vol. 1, no. 4, pp. 282–286, Jan. 2013. [Online].

7. **M**. S. G. Karypis, V. Kumar, and M. Steinbach, "A comparison of docu- ment clustering techniques," in Proc. KDD Workshop Text Mining, 2000, pp. 1–20.

8. **S**. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "Analyzing twitter for social TV: Sentiment extraction for sports," in Proc. 2nd Int. Workshop Future Telev., 2011, pp. 11–18.

9. **A**. W. Ulwick, What Customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services. New York, NY, USA: McGraw-Hill, 2005.

9.1 Text book References

1. **D**eep Learning-Based Approaches for Sentiment Analysis , Basant Agarwal, Richi Nayak, Namita Mittal, Srikanta Patnaik Jan 25 2020