Aastha Patel (Spring 2022)

Analysing the Hot Topic & Favourite Team during Indian Premier League 2022 (2330 words)

Introduction

Cricket is the most popular sport in the world. It is a game of strategy, skill, and endurance and has evolved over the years to become a thrilling spectator sport. There are various tournaments like World Cups, where countries globally participate if qualified; Twenty20 series tournaments where any country hosts the tournament in its home country; also, the Test series. There are many nationally accepted tournaments like the Indian Premier League (IPL) and Pakistan Premier League (PPL). India is a host and rule maker for IPL, allowing the participation of regional teams across India. All International players are invited and auctioned for the teams. The Indian Premier League (IPL) is a highly contentious league on a global scale. It is organized in India by The Board of Control for Cricket in India (BCCI).

Cricket is a particularly well-liked sport in India, where people are usually interested in cricket updates and ongoing events related to it. The only media that updates people instantly about the events is social media such as Instagram, Twitter, and Facebook, where people are also free to express their support and opinions regarding any action. Currently, Twitter is the most well-liked social media tool for getting the latest news. On Twitter, users from all around the world provide updates on news or current events. Because people tend to tweet about things that affect them and their surroundings, analyzing tweets across a period can provide excellent insights into what occurred during that time.

The topics were extracted from the corpus of tweets, which were gathered from each of the IPL teams for identifying the hot topics of IPL. The topic modelling approach used was LDA with TF-IDF. “The main topics are good to understand the overall issues in the document sets and they could form good summaries” [5]. Researchers have used topic modelling extensively in a variety of fields, including sports like Football News where they proposed their study using the LDA model and ended up having overlapped topics clusters that have similarities among the topics [2], and document summarization is done with the same method, as they proposed “to reduce the influence of noise terms in high-weight topics” for future work [3]. As a result, in this study, TF-IDF is used to filter highly used unnecessary terms in all documents and to create unique topics with the least similarities.

Furthermore, sentimental analysis for the other part is carried out to determine the IPL favorite team among the qualified teams.

Research Question

What are the hot topics during the entire Indian Premier League (IPL) 2022 among the participating teams? Which team was the personal favourite of the major public?

Methodology

Data:

Twitter helps with an API to gather the tweets. This analysis includes the use of Twitter API v2 and Tweepy library for data collection. APIs stands for Application Programming Interfaces, are methods that enable two software components to share data with one another using a set of definitions and protocols. Tweepy library is a simple Python library to access Twitter API. Tweets were collected from the Twitter profiles of the teams of the Indian Premier League. These include @mipaltan @ChennaiIPL @DelhiCapitals @gujarat_titans @KKRiders @LucknowIPL @PunjabKingsIPL

@rajasthanroyals @RCBTweets @SunRisers. Since the IPL 2022 season began on March 26 and ended on May 29, the data covered the period from March 1 to May 31. In order to gather tweets that were exciting, encouraging, and in dispute with one another, data was collected from a few days before the start of the IPL 2022 until a few days after its conclusion. Also, the keywords that were used for the data collection were case-sensitive. There were 15000 samples of data points in all. The attributes of data points are Users, Datetime, and Tweets. The team names and the number of tweets gathered are mentioned in Table 1:

Team Names	**Number of tweets**
@gujarat_titans	1568
@RCBTweets	1520
@mipaltan	166
@ChennaiIPL	1667
@DelhiCapitals	1381
@KKRiders	2194
@LucknowIPL	1215
@PunjabKingsIPL	461
@rajasthanroyals	2047
@SunRisers	1633
Total	**13852**

Table 1: Team Name and No. of tweets

Data Extraction for favorite team:

For sentiment analysis, the data was scrapped between four qualified teams, from May 24 to May 29, 2022. The data covered the six days on which the semi-final matches were played between those teams, to get more dense tweets. The samples were gathered using the text search query method of the Python ‘snscrape’ module. It also helped to filter tweets in a certain DateTime range with the maximum required samples.

Here, the keywords were the input parameter in the TwitterSearchScraper method and are not case-sensitive. The keywords for each team are the most used words by a team, and those are also widely used slogans for promoting them. For example, “aavade” (pronunciation: aa-vaa-dey) is a slogan for the team Gujarat Titans. These keywords were got from word cloud, made to get the hot topic of each team. The teams and their keyword used for sentimental analysis is shown in Table 2:

Team Names	**Hashtags**
@gujarat_titans	#aavade
@RCBTweets	#playbold
@LucknowIPL	#LucknowSuperGiants
@rajasthanroyals	#hallabol

Table 2: Team Names and their Hashtags

Analysis:

In this study, Topic Modelling using Latent Dirichlet Allocation (LDA) with Term Frequency Inverse Document Frequency (TF-IDF) is used for identifying the common topics during the IPL 2022. Topic Modelling helps to get abstract topics that occur in a collection of tweet texts (texts are referred to as documents in topic modelling) using a probabilistic model. It is highly used as a tool for mining text to reveal semantic structures within a body of text.

The prior process of data cleaning was done by removing emoticons, @mention, symbols, non-ASCII characters, hyperlinks, special Twitter characters, and numbers. Preprocessing of the data is done firstly by tokenizing, where the content is tokenized by splitting tweet contents into sentences. Furthermore, it splits sentences into words, then lowercase those sentences and removes punctuation. Additionally, all stop words and words with fewer than three letters are eliminated. Afterwards, words are stemmed to their root and lemmatized, where the third person is transformed to the first and future and past tense verbs are changed to the present tense.

A Dictionary of processed words was created which contained the word and the amount of times a word appears in the training set by using the Corpora Dictionary function from the Gensim library. The filtration of the tokens in the dictionary was then carried out by the following conditions:

- First, the tokens were removed if they appeared in less than 15 documents.
- Secondly, the remaining tokens from the dictionary were removed if they appeared in more than 0.5 documents, which meant a particular fraction of the total corpus size.
- At last, keep the first 100k of the tokens from the remaining tokens in the dictionary.

The model was created using 'models.TfidfModel' on the bag of words and Tfidf-corpus was generated by applying a transformation to the entire corpus. The TF-IDF is frequently used to analyze the connections between every word in a group of documents in the fields of text mining and information retrieval. They are specifically used to recognize core words (keywords) in texts, set up search rankings, measure levels of document similarity, and more.

The TF in TF-IDF means the frequency of words in documents. The higher the TF value of a word, the more is its importance. As DF implies the number of times a particular word appears in all the documents. It checks the presence of the word in multiple documents. Unlike TF, a high DF value makes a word less important because it is commonly present in all documents. IDF is an inverse of the DF, used to measure the importance of words in a collection of documents. The higher the IDF value, the more important it is [5].

The result of the topic model is visualized using PyLDAvis model from the Gensim library in Python. It allows us to investigate how topics and terms relate to one another, and the LDA model is better understood as a result. Two panels include an intensity graph in 'pyLDAvis' and a distribution map for each topic. The most intense term (topic) is displayed largest in the intensity graph.

Word clouds are used to show how each topic's data is visualized. A word cloud is a visual representation made up of words from specific text data. Word clouds show how frequently certain words appear in a group of texts. Each word's size reflects how important it is, therefore, if a word appears more often in a topic, then the word size would be bigger. Word clouds were made for each topic. These visualizations were made using the 'wordcloud' library in python.

Analyzing Favourite team:

For analyzing the public's favourite team from four qualified teams for the semi-finals. Sentimental analysis was carried out to get the most loved and supported team using the 'empath' library in python. Data scrapped separately for four teams was cleaned by removing @mentions, symbols and hyperlinks using 'cleantext' library in python.

Empath uses categories, upon which it describes the text belonging to which category by how much portion. It can also be normalized. There are a lot of categories with all the emotions. Some categories were chosen manually for some emotions suited for this analysis. These were further divided into three sentiments, that is "positive", "neutral" and "negative". The samples (tweets) for each team were checked through the chosen categories and sentiments were assigned accordingly. The categories differentiated are shown below:

Positive Categories	Negative Categories
Cheerfulness, Pride, Superhero, Surprise, Leader, Heroic, Celebration, Love, Strength, Power, Achievement, positive_emotion, winner	Aggression, weakness, fear, anger, pain, negative_emotion, loser

Table 3: Distinguish the Positive and Negative Categories

Accordingly, positive, neutral, and negative categorized tweets are then counted and visualized using a bar chart plot of all four qualified teams.

Result

The results are represented, after running LDA model along with TF-IDF by setting parameter of total topics to 6, all the six topics were represented by PyLDAvis model, provided by Gensim library. These topics are visualized in the inter-topic distance map as shown below.

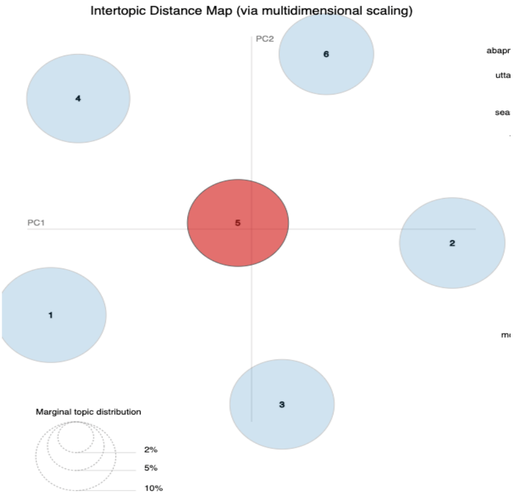

Figure1: Inter-Topic Distance Map

According to the figure 1, all topics are independently clustered. That shows the distribution and frequency of words in the topic are unique.

Top 30 noticeable words in the corpus are shown below in the figure 2. Noticeably, 'TataIPL' was the most frequently appearing term in the corpus, as Tata was sponsoring IPL 2022.

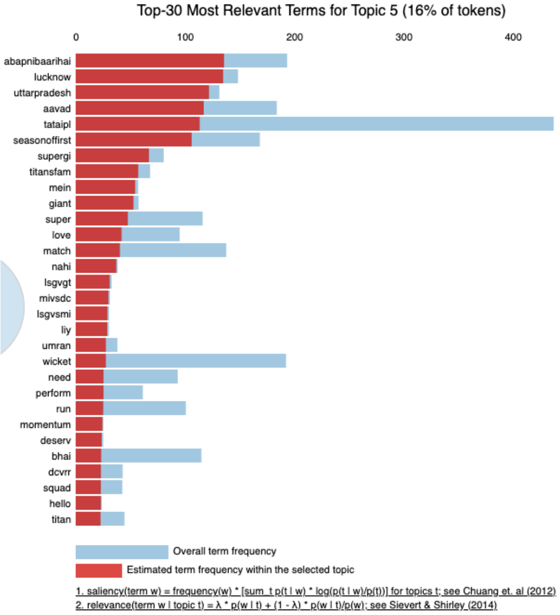

Figure2: Top-30 Most Relevant Terms for each Topic

Word cloud is the best visualization when it comes to showing the importance or frequency of each term in a group of terms. The importance of each term is measured by its size.

Topic 0: Royal Challengers Bangalore Vs Mumbai Indians

The hashtags for Royal Challengers Bangalore, ‘playbold,’ ‘royal,’ and ‘cbrcb,’ are the most often used words in this discussion. Other hashtags include ‘armi,’ ‘mission,’ and ‘dedicates Mumbai Indians.’ However, the terms ‘mission’ and ‘playbold’ outweigh all others.

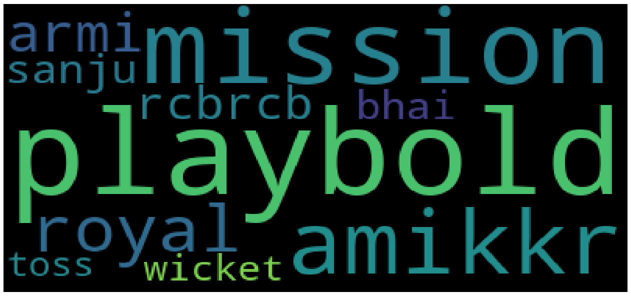

Figure 3: Topic 0 RCB Vs MI

Topic 1: Rajasthan Royals Vs Punjab Kings

The 'hallabol' and 'royalsfamili' are very much used words in this topic. They are used for the Rajasthan Royals team. Moreover, 'saddapunjab' and 'punjabk' are also popular in this topic. As a result, the hashtags 'hallabol' and 'saddapunjab,' however, predominate above all others.

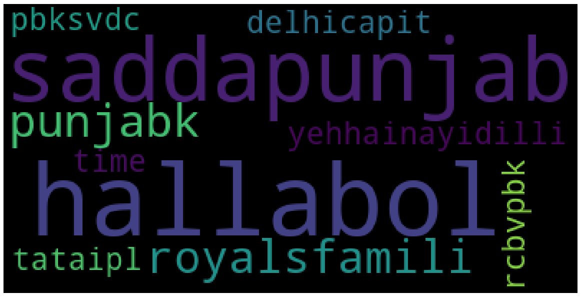

Figure 4 Topic 1 RR Vs PBK

Topic 2: The talks of Chennai Super Kings

'whistlgame', 'whistlepodu', 'yellow', and 'super', are the most often used words in this topic. These hashtags are for Chennai Super Kings. Basically, in this topic, it mostly talks about the Chennai Super Kings (CSK).

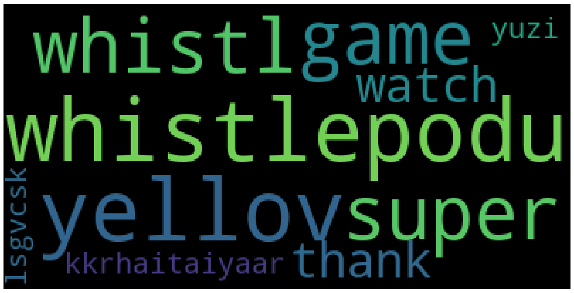

Figure 5 Topic 2 CSK Family

Topic 3: The talks of SunRisers Hyderabad

In this topic, the most used terms are 'readytorise', 'orangearmi', 'riser', which are the hashtags for SunRisers Hyderabad.

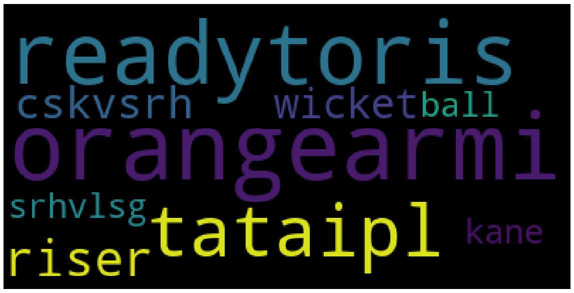

Figure 6 Topic 3 SRH Fans

Topic 4: The talks of new IPL Teams

This topic is named as New IPL Teams as, the most used terms are 'abapnibaarihai', 'uttarpradesh', 'supergi', which are the hashtags for Lucknow Super Giants. Moreover, 'aavad', 'titanfam', 'seasonoffirst' are the tags for Gujarat Titans. Both the teams are new one in IPL 2022

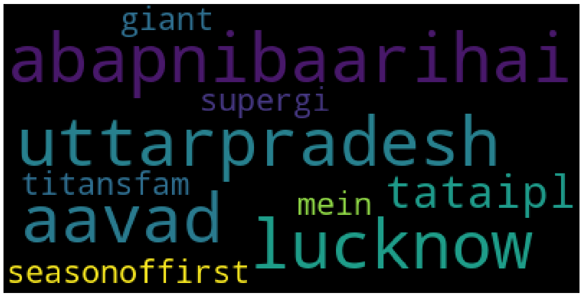

Figure 7: Topic 4 New Teams

Topic 5: The talks of Kolkata Knight Riders

The hashtags 'kkrhaitaiyaar,' 'knigtsinact,' 'kkrvgt,' and 'kkrvmi' appear most frequently in this topic, indicating that tweets were written to motivate the Kolkata Knight Riders.

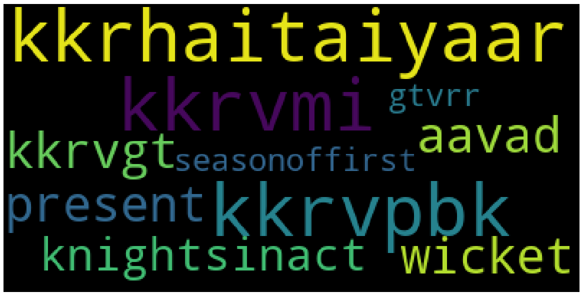

Figure 8 Topic 5 KKR Supporters

These were the hot topics among the Indian Premiere League 2022 teams. Many tweets related to these topics were posted on Twitter to energize the teams. Challenging tweets were also posted between them.

Result of Sentimental Analysis:

The sentimental analysis was conducted on the dataset generated using keywords such as '#hallabol', '#LucknowSuperGiants', '#aavade', and '#playbold'. These are the hashtags of four qualified teams such as Rajasthan Royals, Lucknow Super Giants, Gujarat Titans, Royal Challengers Bangalore respectively. These four hashtags are also popular in the above analysis. The result of the sentiments of the audience is represented by a bar graph with the three sentiments shown in the figure below.

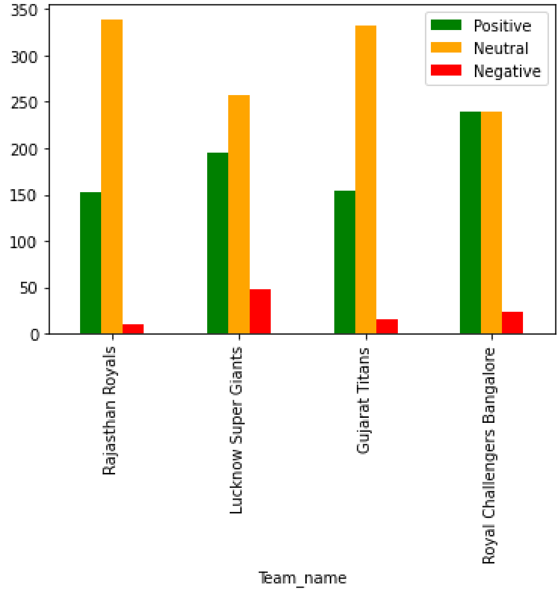

Figure 9: Representation of Sentiment Analysis

Based on the above depiction, Rajasthan Royals has the least support out of the other three teams. In contrast, Royal Challengers Bangalore had the most fans of any of the four teams in the Indian Premier League. Therefore, Royal Challengers Bangalore was the team that the audience favored the most.

Conclusion

In brief, six independent topics were found among the discussions of IPL 2022 teams. Royal Challengers Bangalore was the most supported team of the four qualified teams in the Indian Premier League 2022. Hashtags (keywords) are an important factor of this analysis; but, in some languages, such as #CSK, it may overlook those words because they are abbreviations, which could be more problematic for the study. Many of the tweet samples were in different languages like Hindi, Gujarati, Marathi and so on. As a result, non-English words are eliminated during the data cleaning process, which creates a lack of information in the data.

Reference:

1. Satoshi Sekine, Chikashi Nobata. (2003). A survey for multi-document summarization. In Proceedings of the HLT-NAACL 03 on Text summarization workshop. Association for Computational Linguistics, USA, Volume 5, pp. 65–72. (2003).
   Doi: 10.3115/1119467.1119476

2. F. Hidayatullah, E. C. Pembrani, W. Kurniawan, G. Akbar and R. Pranata. 2018. "Twitter Topic Modeling on Football News". 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 467-471 (2018).
   Doi: 10.1109/CCOMS.2018.8463231

3. J. Bian, Z. Jiang, Q. Chen. 2014. "Research on Multi document Summarization Based on LDA Topic Model". Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics, pp. 113-116 (2014).
   Doi: 10.1109/IHMSC.2014.130

4. L. Zheng, K. Han. "Multi-Topic Distribution Model for Topic Discovery in Twitter". 2013. IEEE Seventh International Conference on Semantic Computing, pp. 420-425 (2013).
   Doi: 10.1109/ICSC.2013.81

5. Kim, SW., Gil, JM. (2019). Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30. (August 2019).
   Doi: 10.1186/s13673-019-0192-7