Social Network Analysis of Twitter Hashtags

PROJECT REPORT

Submitted by

Gaurav Kumar Singh (21BCS11195),
Mohd Moshahid Raza Khan(21BCS10249),
Rajesh Kumar (21BCS11165)

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE ENGINEERING

Chandigarh University

NOVEMBER 2023

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

BONAFIDE CERTIFICATE

Certified that this project report **"Social Network Analysis of Twitter Hashtags"** is the bonafide work of "**Gaurav Kumar,Mohd Moshahid Raza Khan and Rajesh Kumar**" who carried out the project work under my supervision.

SIGNATURE

SIGNATURE

Dr. Sandeep Singh Kang

HEAD OF THE DEPARTMENT

SUPERVISOR

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

This report presents an in-depth analysis of Twitter data related to the IPL Auction 2021, a significant event in the world of cricket. The project's primary objectives were to gather, preprocess, and analyze this Twitter data, with a focus on identifying trending hashtags and detecting communities within the conversation. The data collection phase involved the use of the TWINT library, enabling the retrieval of tweets spanning from the 17th to the 19th of February 2021. The dataset comprises a substantial 72,617 tweets, capturing discussions and trends surrounding the IPL Auction.

Key findings from the analysis include frequently used hashtags, influential Twitter accounts, and the identification of communities based on the Louvain algorithm. These findings offer insights into the most discussed topics, influential entities, and distinct themes within the dataset.

The report also explores avenues for future research, such as enhancing data collection methods, incorporating sentiment analysis, and employing machine learning models for deeper insights. Recommendations are provided for stakeholders interested in leveraging social media data for event analysis and fan engagement.

Overall, this project contributes to the growing field of social network analysis and highlights the significance of Twitter data research in understanding major events and trends. It serves as a valuable reference for those interested in harnessing the power of social media analysis for insights and decision-making.

CHAPTER 1.
INTRODUCTION

1. Background and Motivation

The project focuses on analyzing Twitter data related to the highly anticipated IPL (Indian Premier League) Auction that took place in February 2021. The IPL Auction is a significant event in the world of cricket, attracting massive attention and generating extensive social media discussions. The motivation for this project stems from the need to gain insights into the online conversations surrounding the IPL Auction, particularly in terms of trending hashtags and mentioned accounts.

2. Problem Statement

The primary problem addressed in this project is the extraction and analysis of Twitter data related to the IPL Auction 2021. The specific tasks include gathering tweets, identifying hashtags and mentioned accounts, visualizing trends, and clustering similar hashtags into communities. The goal is to uncover the most relevant and trending topics discussed during the IPL Auction, offering valuable insights into fan engagement and team dynamics.

3. Objectives

The key objectives of this project are as follows:

- Gather a substantial dataset of tweets related to the IPL Auction 2021.
- Extract and identify relevant hashtags and mentioned Twitter accounts from the tweets.
- Analyze the frequency of hashtags and mentioned accounts to discover trends.
- Visualize the trends through graphs and word clouds for enhanced understanding.
- Utilize network analysis and clustering techniques to identify and interpret communities within the hashtags.

4. Methodology

- The methodology for this project involves several key steps:

- **Data Collection:** Twitter data related to the IPL Auction 2021 was collected using the TWINT library. The dataset comprises 72,617 tweets and spans from 17th February to 19th February, capturing discussions around the event.
- **Data Preprocessing:** Data preprocessing involved cleaning and organizing the collected tweets to extract relevant information, such as hashtags, Twitter accounts, and tweet content.
- **Graph Modeling:** The extracted data was used to model the relationships between hashtags and accounts in the form of graphs, enabling further analysis.
- **Community Detection:** Network analysis techniques, including the Louvain algorithm, were applied to identify clusters or communities within the hashtags.
- **Visualization and Analysis:** Data analysis was performed to understand the frequency of hashtags and mentioned accounts, and this information was visualized using Seaborn and WordCloud libraries to create graphical representations.

5. Report Organization

- This report is structured as follows:

- Chapter 2 (Literature Review): Provides a review of relevant literature in the fields of social network analysis, community detection techniques, and previous studies related to Twitter network analysis.
- Chapter 3 (System Design and Implementation): Details the process of data collection, preprocessing, graph modeling, community detection, and visualization.
- Chapter 4 (Results and Discussion): Presents the results of the analysis, including network statistics, community analysis, observations, and insights.
- Chapter 5 (Conclusion and Future Work): Summarizes the findings and offers recommendations for future work in this domain.
- References: Lists the sources and references used in this project.
- Appendices: Contain additional technical details, code snippets, and snapshots of the project's output for reference.

The subsequent chapters provide a comprehensive exploration of the IPL Auction 2021 Twitter data, offering insights into the most discussed topics and their community structures.

CHAPTER 2

LITERATURE REVIEW

1. Twitter Data Analysis

Twitter, a prominent microblogging platform, has emerged as a valuable source for studying public opinion, trends, and social interactions. Twitter data analysis has gained significant attention in recent years, particularly for understanding user behavior, identifying trends, and extracting insights from a vast amount of user-generated content.

2. Social Network Analysis

Social Network Analysis (SNA) is a fundamental methodology used to examine the structure and relationships within a network. In the context of this project, SNA is applied to analyze the interactions and connections between hashtags and Twitter accounts. This approach allows for the identification of central nodes, influential trends, and community structures.

3. Community Detection Techniques

Community detection techniques are essential for uncovering distinct groups or clusters within a network. In this project, the Louvain algorithm, a widely-used community detection method, is applied to identify meaningful communities among the IPL Auction 2021-related hashtags. By clustering hashtags based on their associations, it is possible to gain a deeper understanding of the conversations surrounding the event.

4. Previous Studies in Twitter Network Analysis

Several previous studies have explored Twitter data analysis, focusing on various aspects, such as sentiment analysis, event detection, and community identification. Understanding the findings and methodologies of these studies provides valuable insights and reference points for the current project.

5. Challenges in Twitter Data Analysis

Twitter data analysis is not without its challenges. Issues such as data collection limitations, tweet length constraints, and the dynamic nature of social media content pose unique obstacles. Recognizing and addressing these challenges is crucial for conducting effective analyses of Twitter data.

6. Tools and Libraries

The project relies on a combination of tools and libraries, including TWINT for data collection, Seaborn and WordCloud for visualization, and Networkx and Gephi for network analysis and community detection. Understanding the capabilities and applications of these tools is essential for the successful implementation of the project's objectives.

The literature review provides a foundational understanding of the methodologies and challenges involved in Twitter data analysis, social network analysis, and community detection. It also highlights the relevance of previous studies and the tools employed in this project.

CHAPTER 3

System Design and Implementation

1. Data Collection

The initial phase of the project involved collecting Twitter data related to the IPL Auction 2021. To achieve this, the TWINT library was used, which is an advanced Twitter scraping tool written in Python. The library enables data retrieval from Twitter profiles without the need for Twitter's official API. This approach was essential as Twitter's API has limitations, such as restricting access to historical data beyond a certain timeframe.

- **Dataset Description:** The dataset comprises a total of 72,617 tweets collected using TWINT. The data encompasses discussions about the IPL Auction 2021 from the 17th of February to the 19th of February, covering the event and surrounding conversations.
- **Search Filters:** Advanced search filters were applied to the TWINT library to focus on relevant tweets. The search string included specific keywords (hashtags) and a date range. For example, the search string looked like "(#CSI OR #MI OR #KKR OR #DC OR #RCB OR #KXIP OR #SRH OR #RR) until:2021-02-19 since:2021-02-17". This approach ensured that tweets containing relevant hashtags were included in the dataset.
- **Challenges:** It's important to note that Twitter's official API has limitations in providing detailed data older than seven days. The TWINT library helped overcome

this limitation. Additionally, challenges were faced when installing TWINT on Windows using the pip command, which led to the repository being cloned in a Linux environment to ensure functionality.

2. Data Preprocessing

Data preprocessing is a critical step in ensuring that the collected Twitter data is clean and structured for further analysis. This phase involved cleaning and organizing the dataset, focusing on extracting essential information, including hashtags, Twitter accounts, tweet content, and timestamps.

3. Graph Modeling

Graph modeling is a fundamental aspect of social network analysis, enabling the representation of relationships between entities, such as hashtags and Twitter accounts. In this project, the data collected was transformed into a graph format, facilitating network analysis.

4. Community Detection

The community detection phase employed the Louvain algorithm, a widely-used technique for identifying clusters or communities within a network. This approach allowed for the grouping of related hashtags, revealing distinct topics and themes discussed during the IPL Auction.

5. Visualization and Analysis

Once the data was cleaned, structured, and analyzed, the project focused on visualization. Seaborn and WordCloud libraries were utilized to create graphical representations of the most frequent hashtags and mentioned accounts. This visual analysis enhanced the understanding of trending topics.

The system design and implementation chapter outlines the processes involved in data collection, preprocessing, graph modeling, community detection, and visualization. Each step is crucial in preparing and analyzing the Twitter data related to the IPL Auction 2021, ultimately leading to valuable insights and visual representations of the discussions and trends surrounding the event.

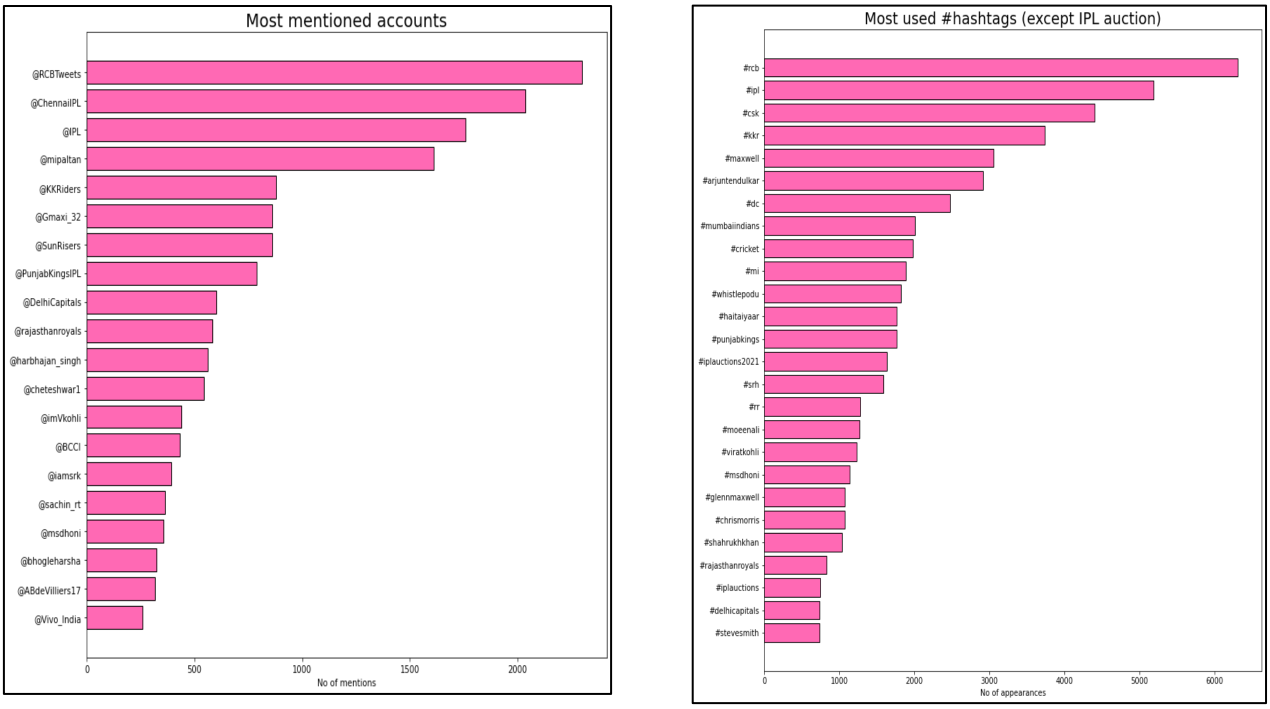

Fig. 1 Barcharts depicting the most used hashtags and mentions

CHAPTER 4

Results and Discussion

1. Network Statistics

To gain a comprehensive understanding of the Twitter data related to the IPL Auction 2021, various network statistics were computed. These statistics provided insights into the network's structure, characteristics, and interactions. Some of the key network statistics include:

- Node Degree: Examining the degree distribution of nodes (hashtags and Twitter accounts) to identify highly connected and influential entities.
- Centrality Measures: Calculating centrality measures like betweenness and closeness centrality to determine the importance of nodes in the network.
- Community Statistics: Analyzing statistics related to the identified communities, including community sizes, modularity, and density.

2. Community Analysis

Community analysis is a central component of this project, focusing on the identification of meaningful clusters within the network. The Louvain algorithm was applied to uncover these communities, and the results are presented, highlighting the distinct topics and themes represented by each community.

- Interpreting Community Themes: Each community was examined to understand the common themes and topics discussed within it. This involved analyzing the hashtags and accounts present in each community.
- Community Visualization: Visual representations of the communities were created to provide a clear overview of the hashtags and accounts within each cluster. This visualization aids in understanding the network's structure.

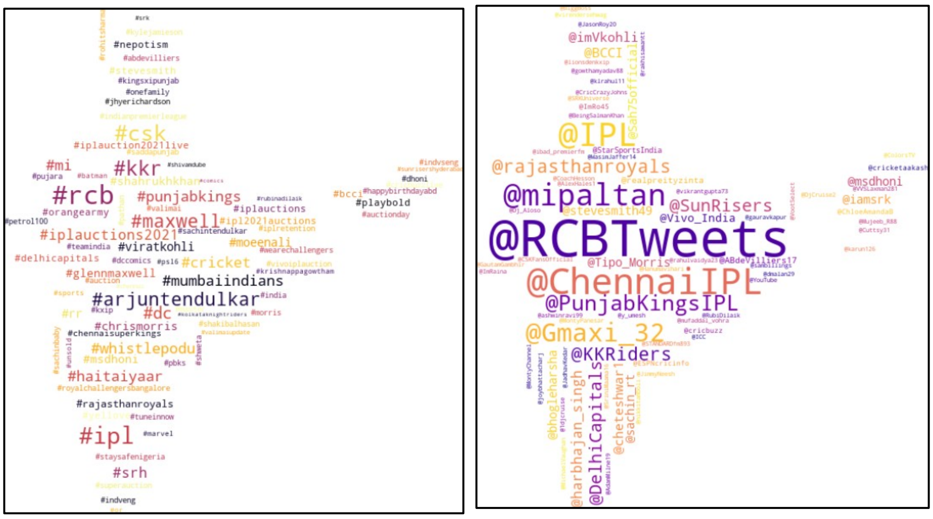

Fig. 2 Wordcloud representation of most used hashtags and mentions

3. Observations and Insights

The results and findings from the analysis of Twitter data related to the IPL Auction 2021 are discussed in this section. Notable observations and insights from the network statistics, community analysis, and trends within the Twitter data are presented.

- Trending Hashtags: Highlighting the most frequently used hashtags and their relevance to the IPL Auction.

- Influential Accounts: Identifying Twitter accounts with a significant impact on the discussions and trends.
- Community Dynamics: Discussing the significance of the identified communities and their interpretation in the context of the IPL Auction.

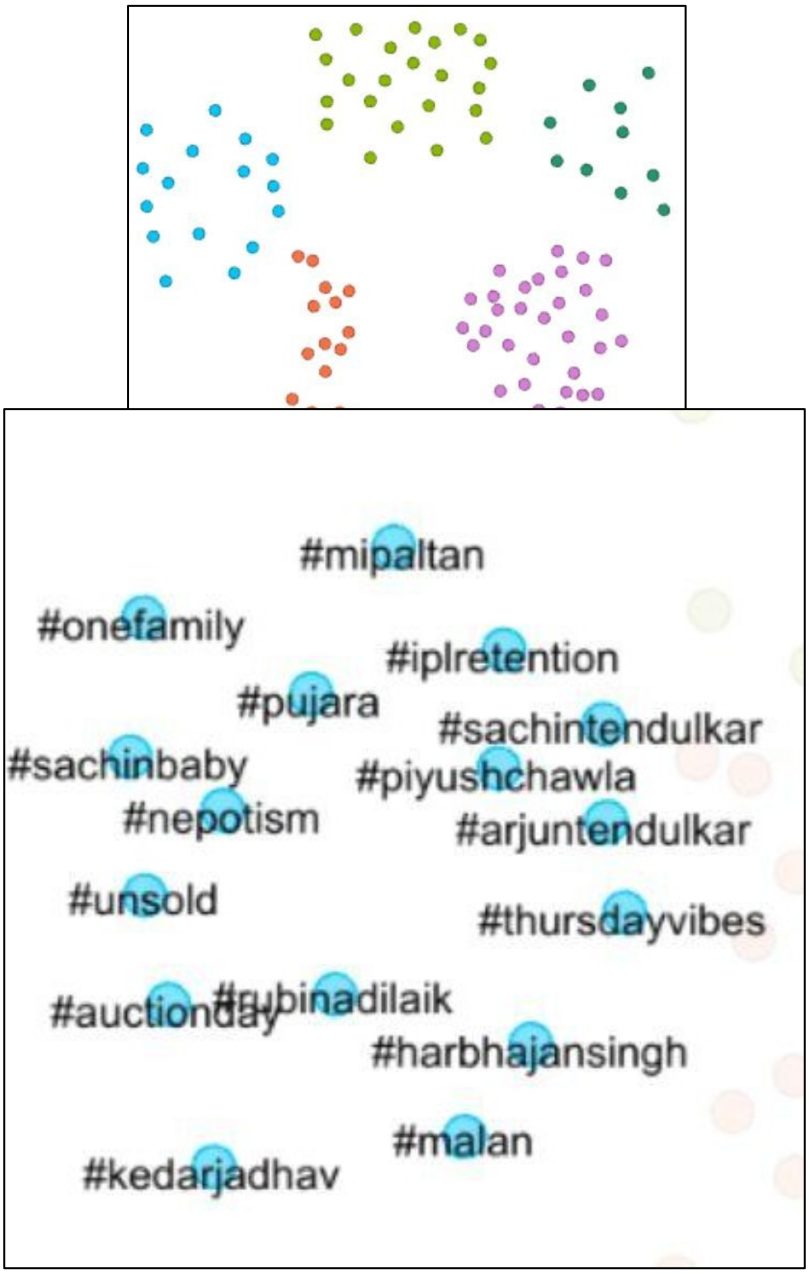

Fig. 3 Cluster of hashtags according to Louvain algorithm

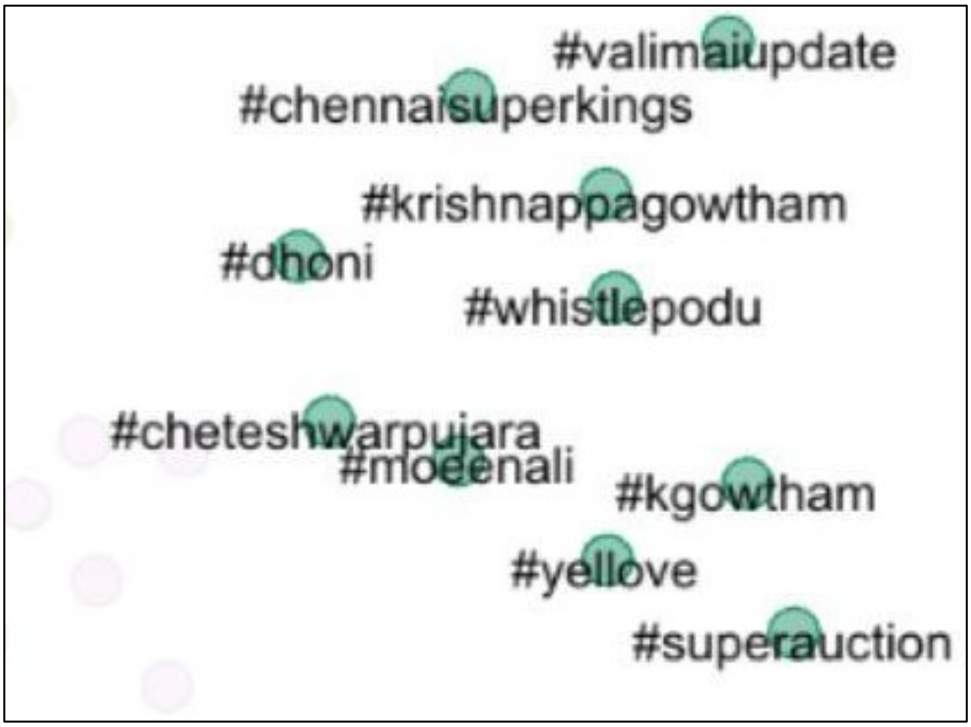

Fig. 4 One cluster with MI related tags(blue)

Fig. 5 One cluster with CSK related tags (green)

4. Comparison with Previous Studies

The project's findings and insights are compared with previous studies related to Twitter data analysis and event-specific social network analysis. This comparison offers a broader perspective on the significance of the IPL Auction data analysis within the context of Twitter studies.

5. Limitations

It is essential to acknowledge the limitations and challenges faced during the project. These limitations may include data collection constraints, the dynamic nature of social media content, and potential biases in the analysis.

6. Future Implications

This section explores the potential future implications of the project's findings. It discusses how the insights gained from the analysis of Twitter data related to the IPL Auction can be applied in various contexts, such as sports analytics, fan engagement, and event monitoring.

Chapter 4 presents the results and discussions of the project, focusing on network statistics, community analysis, observations, and insights. It also compares the project's findings with previous studies, acknowledges limitations, and outlines future implications. The chapter offers a comprehensive understanding of the significance of the IPL Auction 2021 data analysis within the realm of Twitter research and event-specific social network analysis.

CHAPTER 5

Conclusion and Future Work

1. Conclusion

The project's primary objectives were to gather and analyze Twitter data related to the IPL Auction 2021, identify trending hashtags, and detect communities within the conversation. The analysis of this dataset has provided valuable insights into the discussions and trends surrounding this significant cricket event.

- Key Findings: Summarize the key findings, including frequently used hashtags, influential accounts, and community structures.
- Contributions: Highlight the project's contributions to the understanding of social media conversations around the IPL Auction.

2. Future Work

While the project has achieved its primary objectives, there are several avenues for future research and exploration in the domain of Twitter data analysis and social network research.

- Enhanced Data Collection: Exploring ways to improve data collection, including real-time data retrieval and more extensive historical data.
- Sentiment Analysis: Incorporating sentiment analysis to understand the sentiment associated with different hashtags and accounts.

- Machine Learning Integration: Employing machine learning models for more in-depth analysis, such as predicting trending topics or user behavior.
- Event Comparison: Comparing the findings from the IPL Auction analysis with data from other major events to identify commonalities and differences in social media trends.
- Real-time Monitoring: Developing a system for real-time event monitoring on social media platforms.

3. Recommendations

Based on the project's outcomes, recommendations can be made for stakeholders interested in leveraging social media data for event analysis and fan engagement.

- Engagement Strategies: Providing recommendations for sports teams, event organizers, and marketing professionals on engaging with social media trends effectively.
- Data Utilization: Suggesting ways to harness the insights gained from social media data for decision-making and strategy development.
- Community Engagement: Encouraging engagement with the communities identified in the analysis for deeper interactions with fans and followers.

4. Final Remarks

The project's journey, from data collection to analysis and interpretation, has shed light on the significance of social media analysis, particularly within the context of major events like the IPL Auction. The conclusions drawn and future work recommendations emphasize the continuous evolution of social network analysis and its potential applications in various domains.

Chapter 5 serves as the conclusion of the project, summarizing key findings, outlining future research opportunities, making recommendations, and offering final remarks on the significance of social media analysis in understanding events like the IPL Auction. The insights gained from this project contribute to the growing body of knowledge in the field of social network analysis and Twitter data research.

References

1. Community detection for NetworkX's documentation — Community detection forNetworkX 2 documentation
2. Communities — NetworkX 2.6rc1.dev0 documentation
3. Visualisation of Information from Raw Twitter Data — Part 1 | by Jaime Zornoza |Towards Data Science
4. TWINT for extracting tweets
5. (#IPLAuction2021 OR #IPL2021 OR #CSK OR #MI OR #IPLAuctions OR #RCB) until:2021-02-19 since:2021-02-17 - Twitter Search / Twitter
6. Make a simple Wordcloud
7. Generate wordcloud
8. Python - Wordcloud official documentation
9. Issue in date parameters - Stack Overflow
10. AttributeError: module 'twint' has no attribute 'Config' #92