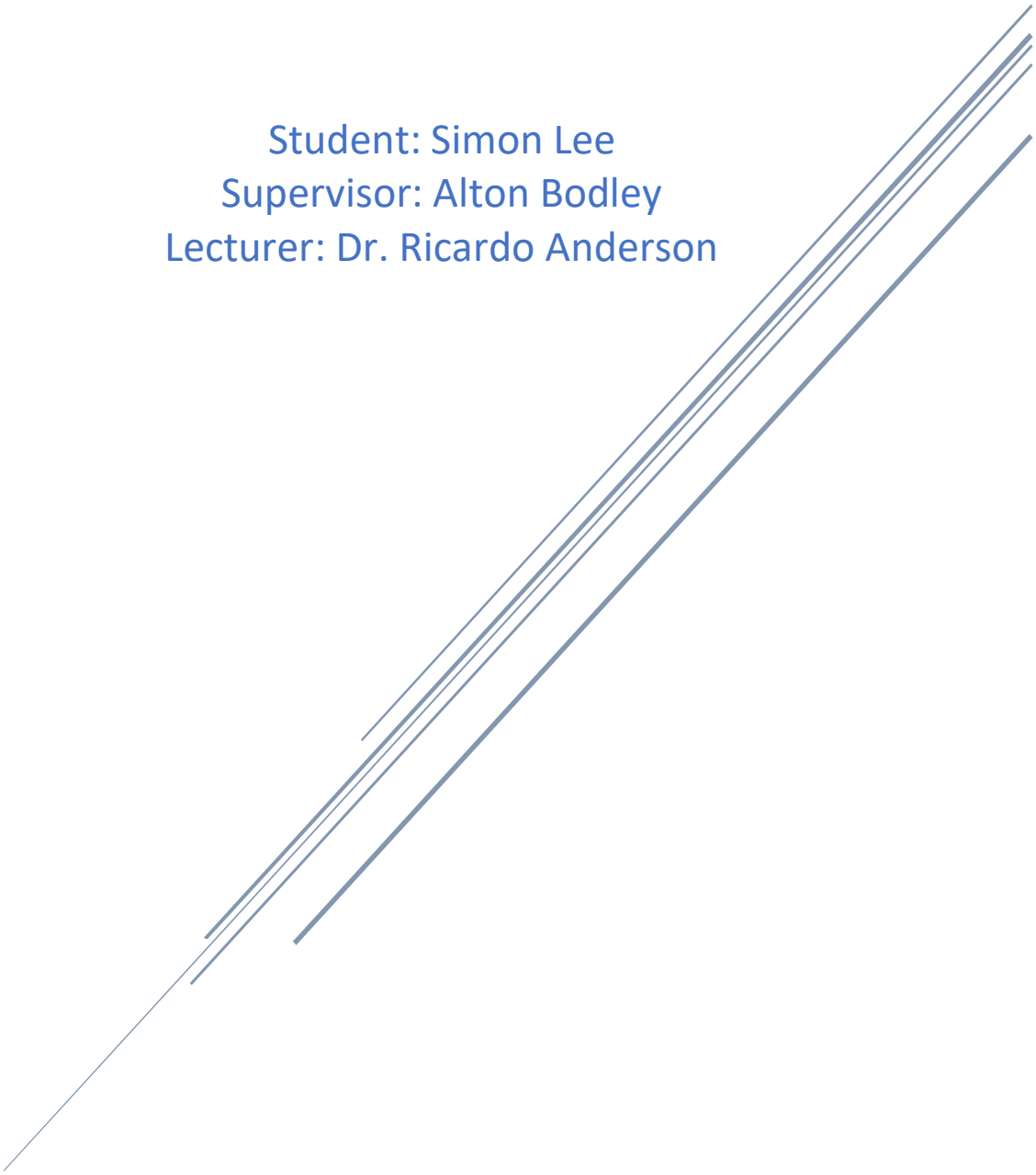
August 18, 2022

TOWARDS ADVANCING T20 CRICKET

TECHNICAL DOCUMENTATION

Student: Simon Lee
Supervisor: Alton Bodley
Lecturer: Dr. Ricardo Anderson



The University of the West Indies, Mona
Applied Data Science Capstone Project

Table of Contents

Background

Cricket is a sport that started in the 16th century, originally lasting up to five (5) days. Dubbed as the gentleman's sport, cricket was played amongst the elite of society and was recognized as a game for the sophisticated. Fast forward, the sport has grown in many ways. The sport is now far more accessible to the general population as is now one of the top three (3) most popular sports in the world. Simply put, the aim of cricket is to make more runs that the opposing team. The game has evolved into new formats, namely ODI Cricket, which allows each team a maximum of fifty (50) over each and T20 Cricket, the newest format of cricket, which allows teams a maximum of twenty (20) overs each.

This shiny new format was not widely accepted by the traditional fan of cricket. However, it has grown to become the most watched format of cricket today. With the increase in viewership and the fast-paced, exciting nature of T20 Cricket, revenues have blown up through television viewership, sponsorship and general population of the sport. Hence it is imperative to ensure fans remain satisfied with the product as well as ensuring that the product continues to improve.

Objectives

The aim of this project is to assist in advancing the T20 format of cricket both performance-wise and business-wise. As such, there are two main objectives for this research.

1. Advance Performance by Building a Team Selection Optimizer

- Use match level data (Team Compositions) and player level data (statistics of each player in the selection pool) in order to develop a model that will generate the optimal team.

The main benefit of this Team Selection Optimizer is that team selection will become easier as well as the fact that bias is taken out of team selection. Hence, players will be giver fair and equal opportunities. An additional benefit is that the performance of the teams should increase as it is expected that the most efficient composition of teams will be playing against each other.

NB: The makeup of a balanced cricket team must be kept intact. That is, the team must include a wicketkeeper and at least five (5) players considered to be bowlers.

2. Improve the Connection with Fans Around the World

- Collect tweets on T20 cricket from around the globe and analyze how receptive people are to T20 cricket.

This will allow the stakeholders in T20 cricket to see how the public's perception of T20 Cricket has changed over time, from inception to now.

T20 Cricket Fan Sentiment Analysis

Introduction

Sports are constantly evolving, whether it be rule changes or even introducing a whole new format. There is no difference when it comes on to cricket. Traditionally, cricket was played over a period of five (5) days, where teams had an unlimited number of overs that they could bat. However, in 1971 One Day International cricket was created. As the name suggests, this new format spanned one day, with each team having fifty (50) overs each to bat. In 2006 stakeholders saw fit for another format which saw the birth of T20 Cricket. This format saw teams battling over four (4) hours with twenty (20) overs to make as many runs as possible. Fast forward more than fifteen years later, T20 cricket has become the most played format of cricket with tournaments being held all over the world all year round. However, some may argue that T20 cricket was not always as popular as it has become. In fact, there was some backlash from the traditional cricket fans when the new format was announced. This research looks to get insight on how the general population's feeling on T20 cricket has developed over time.

Data Collection

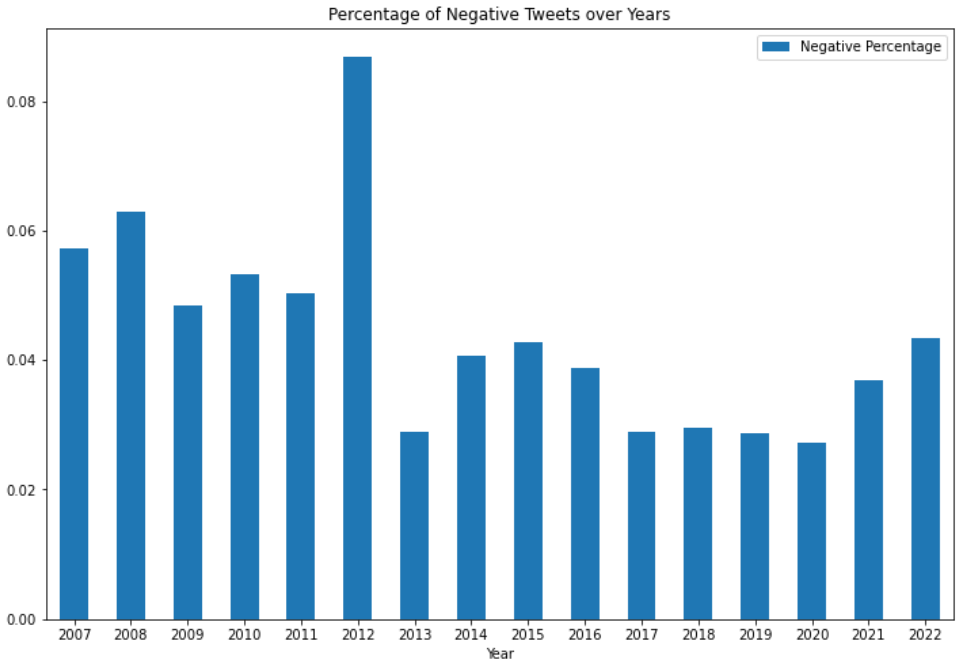
The data used in this research comes from twitter, a social media platform which allows people to share their opinions. In this research, the scscrape package was used, and more specifically, sntwitter to pull historical tweets. Tweets containing "T20 Cricket" made between the years 2007 and 2022 were collected and capped at twenty thousand (20,000) per year. However due to the fact the twitter was started in 2007, that year and the following year did not come close to the 20,000 tweets. Therefore "Twenty Twenty Cricket" and "20 20 Cricket" were searched as alternative names for T20 Cricket for those years to gather more tweets for our dataset.

Data Preparation and Method

The texts of the tweets were passed through a data cleaning function to remove all special characters. Additionally, and each tweet was tagged with their corresponding year. Afterwards, The texts of the tweets were passed through the TextBlob function, and the sentiment of each tweet was attained. Those tweets with a sentiment score of less than -0.2 were labeled as negative, the tweets above 0.2 were labelled as positive, and those in between as neutral. Once each tweet had a sentiment of either positive, negative or neutral assigned to it, the tweets were grouped by year and a count of all the positive and negative tweets were taken. This allowed us to get what percentage of our tweets pulled for each year was positive and negative.

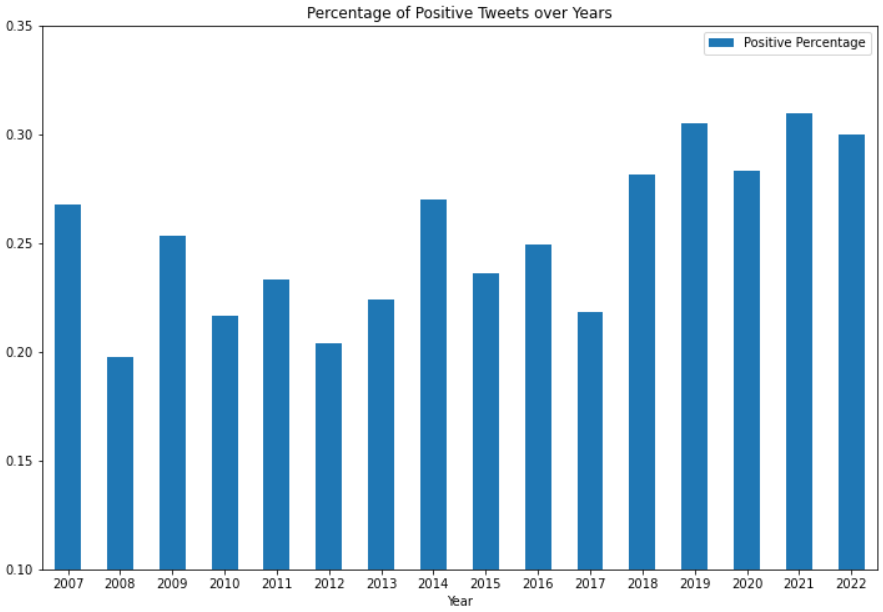
Results

Firstly, examination was done on the percentage of tweets that were negative per year. The bar chart below was generated.



As can be seen from the above chart, the percentage of negative tweets has been trending downwards (about two percentile) and would suggest that the newest format of cricket, T20 Cricket has been more accepted over time and has increased in popularity amongst fans of the sport.

However, it would be insufficient to only analyze negative tweets to come to that conclusion. Hence, the same bar chart was generated for positive tweets. The resulting chart is shown below.



In the figure above, we can see the opposite of the chart depicting negative tweet percentages. This suggests that the positivity around T20 Cricket has trended upwards (roughly 8 percentile), and thus increased over time. It also confirms that though the percentage of negative tweets was trending downwards, the tweets were turning into positive tweets and not neutral tweets. Upon further inspection with the aid of a world cloud seen below,

it may be useful to note that fans think the new format of cricket is exciting as the increase in the number of sixes hit mixed with the fast-paced nature of the game has got fans very interested and upbeat.

Conclusion

Based on the sentiments drawn from the tweets spanning sixteen (16) years from 2007 – 2022, it can be concluded that the popularity of the T20 Cricket format has grown and has done so positively. This can be seen as the percentage of negative tweets per year has trended downwards and the percentage positive tweets per year, upwards. This gives great reason to believe that over time the fans have been more accepting of the new format of cricket. With proposed rule changes to T20 Cricket on the horizon, a further enhancement of the game is due and based on this study, it can be assumed that the fans will take well to it.

Limitations

- The location of the tweets was unattainable and so it was impossible to drill down on specific locations and their general sentiment towards T20 Cricket.
- Twitter started in 2007, one year after T20 Cricket started, and thus data for 2006 was unavailable. It also made collecting the capacity (20,000 tweets) for the years 2007 and 2008 as the social media platform had not grown in popularity yet.

T20 Cricket Team Selection Model

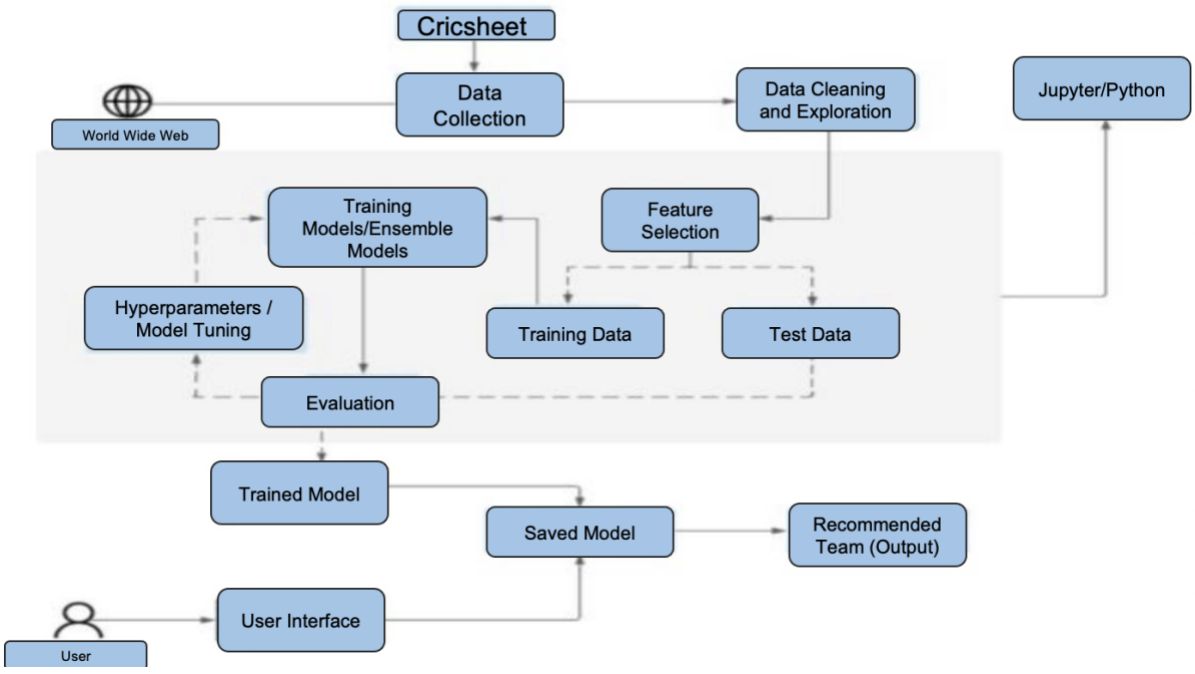
Introduction

Team selection is a constant conundrum in the sporting world. Knowing what combination of players give a team the best chance of winning is almost as critical as how the players perform. Sporting teams ideally want to give themselves the best chance of winning when selecting a team, however due to human bias and lack of insight this may sometimes not be the case. Team selection is therefore critical. However, the decision is that much more important when it comes onto cricket. This is so, as unlike other sports, cricket does not allow for substitutions unless serious injury has occurred, so the team that is selected is the team that you must rely on for the duration of the match.

This project intends to remove human bias as well as optimize the team selection process as best as possible. In doing so, the process will be separated into two parts:

1. Finding the best team composition (how many of each type of player the team should have).
2. Selecting the players that are most fit to occupy these positions given the opposition.

The project followed this process flow.



Data Mining

The data used in this project was collected from CrickSheet (https://cricsheet.org). From this website, both match data as well as player data (which contained player IDs to other websites) were downloaded as csv files. Each match's data contained information on the match such as venue, teams involved, outcome of the match, players involved in the match as well as their "CricSheet" ID and lastly, details on each ball of the match. The ball-by-ball data contained the players involved in the ball (the bowler and the batsmen), how many runs were scored, whether a wicket fell and nature of the dismissal.

Along with the Match level data, a player document was also downloaded from CricSheet, which provided many IDs for each player for different websites. This was leveraged by getting the Cricinfo ID for each player. With this data, player information could be scraped from ESPN Cricinfo (the most used Cricket website), such as their player roles, batting styles and bowling styles. These files were rewritten into various csv files for ball-by-ball data, match data, player data and team composition data which shows how many of each type of player was used per team.

Data Cleaning and Preparation

Once the data was consolidated and stored in the various csv files, a few steps were taken to prepare the data that would be used for modeling and team selection.

- Team Composition
- Categorizing Bowling Styles

Players were placed into one of the following bowling styles:

1. Right Arm Fast
2. Left Arm Fast
3. Right Arm Wrist Spinner
4. Left Arm Wrist Spinner
5. Right Arm Spinner
6. Left Arm Spinner
7. No Bowl (Those players that had no bowling style on ESPN CricInfo)

- Treating Missing Values for Player Roles

For Players that had missing Player Roles, ball by ball data was used to get the players' usage. That is, the number of balls a player batted per match and the number of balls bowled per match were taken into consideration. The following table represents the flow of how player roles were assigned.

Balls Batted per Match	Balls Bowled per Match	Player Role
0	>0	Batter
>0	0	Bowler
>10	>6	Allrounder
>10	<6	Batter
<10	>6	Bowler
>Balls bowled per match		Batter
	>Balls faced per match	Bowler

- Creating Team Composition

From the match data, all IDs of players in each team was available. Using these player IDs, as well as the player roles file generated from scraping ESPN CricInfo, it was possible to get a data set containing the number of each

1. Batting Styles - Left Hand Batter, Right Hand Batter or No Batting Style
2. Bowling Styles - Left/Right Arm Fast, Left/Right Arm Spinner, Left/Right Arm Wrist Spinner, No Bowling Type
3. Player Roles – Opening Batsman, Top Order Batsman, Batter, Batting Allrounder, Allrounder, Bowling Allrounder, Wicketkeeper, Bowler

Note: Each of these categories added to 11 players each.

- Player Statistics

➢ Generate Player Statistics

Using the ball-by-ball data, which shows the event that took place on each delivery, player stats were generated. These statistics included:

1. Batting Average – $Total\ runs\ made\ \div Total\ Dismissals$

Batting average will indicate to us the amount of runs a better is expected to make before he is dismissed.

2. Batting Strike Rate – $Total\ number\ of\ runs\ made\ \div Total\ tNumber\ of\ balls\ faced$

Batting strike rate will allow us to know how quickly a batter scores his runs. This is so as it indicates how many runs are scored per ball (in some cases per 100 balls).

3. Bowling Average - $Total\ runs\ conceded\ \div Total\ tNumber\ of\ wickets\ taken$

This indicates how many runs will be scored off a bowler before he gets a wicket.

4. Economy Rate - $Total\ runs\ conceded\ \div Number\ of\ Overs\ Bowled$

Economy rate is an indicator of how fast batters score off a particular bowler.

➢ Players With No Dismissals or Wickets taken

The players with no wickets or dismissals to their name posed a problem as their bowling/batting average would be ‘infinity’ as the denominator of the formula would be zero. To solve for this problem, the infinity values were replaced with the total number of runs scored by the batter, and in the case of the bowling average, the total number of runs the bowler conceded by the bowler.

- Categorizing and Label Encoding Location of Match

Using the city of where the match was being played, a new column "Country" was created which showed the country/region which the match was played. The following groupings were used:

1. Australia
2. England
3. New Zealand - NZ
4. South Africa - SA
5. Zimbabwe
6. West Indies – WI (Caribbean Islands + Guyana)
7. Sub-Continent (India, Sri Lanka, Bangladesh, Pakistan)
8. United Kingdom – UK (Ireland, Scotland)

The reason England was not included in the category of UK is that England conditions tend to be different to those of Ireland and Scotland. It can also be noted that conditions across all sub-continent countries tend to be similar.

After the cities were categorized into countries/regions, Label Encoding was done to end up with numerical categorical values for modeling.

Modeling and Evaluation

Once the data was cleaned and prepared, it was time to start building a model that would be capable of predicting match outcome from team compositions of the home and away teams and the country/region that the match was being played in. There were three classifier algorithms that was tried and tested. These were Decision Trees, Random Forrest and ExtraTreesClassifier. Before applying these algorithms, the data was transformed using Power Transformer to reduce skewness and give if more of a Gaussian distribution. Additionally, GridSearchCV was used to find the ideal combination of parameters for each classifier. Once this was done it could be noticed that Decision Trees performed at an accuracy of roughly 53% with a clear indication of over-fitting. Meanwhile, Random Forrest performed slightly better giving an accuracy score of around 57%. Ultimately, the ExtraTreesClassifier was used as it gave an accuracy of 61%. This algorithm splits the data in small sub-parts and applies decision trees on the then takes an average of the rules formed to create a final classifier. This method is like that taken in Random Forrest but allows for better accuracy scores as well as a reduction in overfitting of data. Once that model was chosen it was noticeable that it was able to predict away wins (66%) slightly better than home wins (57%).

Deployment

After the model was created, it was time for it to be deployed. The way in which this is carried out is by first entering a list of players from which the user can select their playing 11 from, called the selection pool. Once that list is entered, all possible combinations of that team will be found, and the compositions of those teams will be generated. After which, the legitimate teams will be stored. Please note that a team is noted as legitimate if it has at least five (4) batters and five (5) players capable of bowling (bowlers, allrounders, batting allrounders and bowling allrounders) and one (1) wicketkeeper. This is done to ensure that the team chosen has realistic balance as well as the fact that at least five (5) bowlers must bowl in a match.

The next step is to enter the expected players that will make up the playing 11 of the opposition as well as the country/region the match will take place in. The composition of the team for that specific mix of players will be found. Now, the composition of each of the legitimate possible

teams from the selection pool will be passed into the home team variables and the composition of the expected playing 11 of the opposition will be entered in the away team variables. Once all the possible combinations are exhausted, the composition of players that gives the highest probability of a home win will be chosen.

Now, if there are more than one way to fill in this composition with the players available in the selection pool, each option is checked by taking a sum of the batting index of each possible playing 11 and then selecting the team that has the highest overall batting index (strike rate * average/50). This team is then returned to the user along with its chance of winning.

Please see file Interface.py in the files provided for deployment. To start app, please have streamlit, the player roles csv file, the player statistics csv files and the model downloaded. Ensure the interface.py file points to these csv files on your machine as well as the model. Once that is done, enter "streamlit run Interface.py" in your command prompt and the app should launch.

Limitations

1. The recent form of players is not taken into consideration, which is one additional variable that may affect team selection. With additional time, this would have been a potential variable to investigate.

2. Some players in this data set were not popular enough to have their player batting style, bowling styles and player roles filled out.

Conclusion

In concluding let's analyze how well the project has done in comparison to its initial objectives.

1. Building of a T20 Cricket Team Selector

This objective was successfully completed with the use of using a model that was found to be 61% accurate as well as using an app for user interaction with the aid of streamlit. However, improvements can be made to this team selector by possibly integrating player recent form into the selection process as well as other considerations.

2. T20 Cricket Fan Sentiment

Likewise, this objective was also met, as we were able to understand the change in fan sentiment from 2007 to present. It showed that the general sentiment has trended more positive as the years went on and less negative. However, we were unable to get the location of these tweets which may have added more insight on where the T20 format is not doing so well across the world.

References

Badhan, R., Shesir, M. and Fakir, N., 2018. “Squad Selection For Cricket Team Using Machine Learning Algorithm”.

M. G. Jhanwar and V. Pudi, "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach," in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2016 2016), 2016.

G. D. I. Barr and B. S. Kantor, "A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket", Operational Research Society, vol. 55, no. 12, pp. 1266-1274, December 2004.