*Player Analysis Dashboard*

Virat Kohli – Indian Cricket Batsman

*Akshata Thopte, Kunal Chopra, Riddhi Umap, Siddharth Lakhani, Uthra Venkatakrishnan*

Table of Contents

[**Introduction** 2](#_Toc121174445)

[**Business Objective** 2](#_Toc121174446)

[**Data** 3](#_Toc121174447)

[**Information Quality** 5](#_Toc121174448)

[**Methods and Tools** 5](#_Toc121174449)

[**Data Wrangling Process** 6](#_Toc121174450)

[**Analysis and results** 8](#_Toc121174451)

[**Screenshots of the dashboard** 9](#_Toc121174452)

[**External Materials used** 10](#_Toc121174453)

[**Additional Data and analysis** 11](#_Toc121174454)

## **Introduction**

Text mining is a kind of machine learning that analyzes and retrieves meaning from unstructured text information using Natural Language Processing. This can be used to comprehend the tone of a particular text document, the context of the text, and various terms and phrases related to the key subject or character in the text. Extraction of meaning from social media and web data would aid in solving many business challenges since it has a lot of unexplored information. Text mining is useful for deciphering and extracting meaning from blogs and social media like Facebook, Twitter etc. To gauge the people ’s perspective of him, we collated tweets about him during the world cup 2022 and conducted multiple analyses using those results. We utilized statistics from the well-known sports organization ESPN to get his batting performance. These datasets were also incorporated to provide a 360-degree profile review of Virat Kohli.

## **Business Objective**

This project's main goal is to aid a cricket franchise in choosing the ideal group of players who will have excellent public reputations in addition to strong playing skills. By having a well-liked player on their squad, the club can win more trophies and gain more popularity. As a result, the franchise will be able to profit more from brand sponsorships, marketing, etc. To do this, we created a prototype for an interactive dashboard that portrays the sentiments and playing capacities for the well-known Indian cricket player Virat Kohli across various attributes. This also applies to other players and sports. A client can use this dashboard to determine whether to select a player or not by understanding the trends in the player's performance over time, including across different match formats, years, and public perception.

**Applications of the dashboard**

The interactive dashboard can be utilized by

1. Franchise sports teams to find the best player for their team

2. Using this dashboard, betting organizations can choose how many points and how much money to award certain individuals for betting

3. Product firms can use a player's emotional appeal to determine whether they should select them to represent their brand

4. Sports broadcasting firms can create articles, posts, etc. using this dashboard

5. Using this dashboard, sports fans may keep up with the most recent player information

## **Data**

Data was gathered from two independent sources: Twitter and ESPN

Twitter datasets

Data on tweets, retweets, and other information pertaining to Virat Kohli and his overall performance were retrieved using Twitter APIs. The dataset contains a wide range of attributes, including user ID, location, date of creation, tweet content, source, description, and username, and other. A total of 5000 rows per dataset, for November 2, 6, and 10, 2022, were collected. India played in high-profile matches against England, Bangladesh, and Zimbabwe on these dates.

*Twitter data snippet*

Graphical user interface

Description automatically generated

ESPN datasets

Seven datasets containing the batting stats of Virat Kohli were scraped from ESPN’s official website using excel web import services

* Batting stats across different match formats
* Batting stats across different cricketing countries
* Batting stats across multiple years
* Batting stats across different match results
* Batting stats across multiple tournaments
* Batting stats across multiple batting orders

*ESPN Data Snippet*

*Table

Description automatically generated*

Table

Description automatically generated

ESPN link : <https://www.espncricinfo.com/player/virat-kohli-253802/bowling-batting-stats>

## **Information Quality**

Data Quality Issues:

* The Twitter API spits out a plethora of information and attributes for each tweet
* The text and schematic data retrieved from the API are both unstructured
* The text in the tweets has plenty of noise
* The Twitter API request includes time limits and row limits that restrict the data pull to certain numbers and may result in information loss
* The traceback time is only 7 days, making it challenging to recover historical data for analytics

Solution

Most issues with data quality were resolved by the wrangling procedure utilized.

## **Methods and Tools**

Logo, company name

Description automatically generated

Postman was used to view and validate the API data from Twitter

Excel was used for basic data Wrangling, profiling, and web data pull

Icon

Description automatically generated

Icon

Description automatically generated

Python was used for all the data wrangling and data analyses processes

A picture containing clipart, brick, building material

Description automatically generated

Lucid Chart was utilized to make process flow diagrams for data wrangling process

Microsoft Power BI was used to make the prototype of the interactive dashboard

Logo, icon

Description automatically generated

## **Data Wrangling Process**

Since we are dealing with text data in this study and text data requires a substantial amount of data cleaning before it can be used for any sort of analyses, it was vital for a structured data wrangling process in this situation. The steps we took to sanitize the tweets data are listed below

The Twitter Data was pulled using Twitter API and was imported into Jupyter notebook as a dataframe

Icon

Description automatically generated

**Data Formatting**

The dataframe consists of columns that has Json objects in the form of string

* User Defined functions were created to convert these String objects to JSON
* The converted JSON objects were flattened into columns in the same Dataframe

Icon

Description automatically generated with medium confidence

**Data Cleaning**

To eliminate the noise from the data, text cleaning involves implementing many distinctive and specialized methods. In this situation, we employed the fundamental text cleaning techniques

* Removing Duplicates from tweet data
* **Elimination of punctuations:** Tweets involve a lot of punctions which do not contribute much from an analysis standpoint and therefore were removed

Punctuations were removed from the text using regular expressions

* **Tokenization of data:** Tokenizing is a process of dividing a corpus into its basic meaningful entities. This is usually text, but it might also be hashtags, emojis, or other symbols. Tweets are especially intriguing since various hashtags, emoticons, and other interesting symbols have diverse meanings
* **Stopword Elimination:** Removes words such as the, a, and, etc. which occurs a lot in a text but doesn’t add any value in extracting the context of the text
* **Stemming and Lemmatization:** Stemming extracts the roots of the word. Lemmatization takes the root of the word considering the context in which the word falls

Icon

Description automatically generated

**Data Enrichment**

Included variables like

* Tweets stemmed

Icon

Description automatically generated

Each one of it will aid in specific analyses

Included a separate dataset that contains the word and its frequency to understand the most occurred word in the entire tweet dataset

Unknown columns and the columns that were not necessary for the analyses were dropped

**Missing Value Treatment**

Columns that had more than 90% null values were removed

Icon

Description automatically generated

For ESPN datasets following were the data wrangling steps implemented

1. Changed the column names for better interpretability
2. Dropped rows were the number of matches played were less than 10
3. Fixed the datatypes: converted all object datatypes to int or float to help in carrying out arithmetic calculations
4. Removed additional values from the column values using string replace functions

**Process Flow Diagram**

**Graphical user interface, text, application

Description automatically generated**

## **Analysis and results**

We performed basic exploratory analyses to understand the different kind of data available and majorly performed sentiment analysis on the tweets data. This analysis was carried out using the python package named ***textblob***

Textblob has two parameters ***Polarity*** and ***Subjectivity*** based on which sentiment is provided to a text as positive, negative, or neutral. Polarity values lie between -1 and 1. If the values are closer to 1, then it means it’s a positive sentence and if a value lies closer to -1, then it’s a negative sentence. Subjectivity is a float value that lies between 0 and 1. Each text was passed into a for loop expression based on these values to obtain a sentiment tag (Positive, negative, or neutral). We created an interactive dashboard that can be used to comprehend a player's value at various levels based on sentiment data and player performance data.

Metrics and Dimensions of the Dashboard

|  |  |  |
| --- | --- | --- |
| Metrics | Dimensions | Types of Visuals Created |
| Highest runs | Match formats – ODI, T20, Test | Tiles/Cards |
| # of 100s  # of 50s  # of 6s  # of 4s  # of 0s | Across different Match results   * Win * Lose * Draw * Tied | Donut Chart |
| # of Matches  # of Runs  Strike Rate | * Year * Batting Position (1-7) * Played Against (Country Names) | Combo charts (Line and Bar) |
| Total Positive Sentiment  Total Negative Sentiment  Total Neutral Sentiment | Date  User Location | Heat Map,  Bar Graphs |
| Total Tweets, Total Retweets, Total favorite count | Date  User Location | Cards/Tiles |
| Words Associated with Virat Kohli | Date  User Location | Word Cloud |
| Top 10 Words associated with Virat Kohli | Date  User Location | Bar Graph |

## **Screenshots of the dashboard**

**Graphical user interface, application, email, website

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

## **External Materials used**

* [https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using- -textblob/#:~:text=Polarity%20is%20float%20which%20lies,of%20%5B0%2C1%5D](https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-%20-textblob/#:~:text=Polarity%20is%20float%20which%20lies,of%20%5B0%2C1%5D) - This was utilized to understand polarity and subjectivity
* <https://www.kaggle.com/code/paoloripamonti/twitter-sentiment-analysis> - sentiment analysis python
* <https://www.kaggle.com/code/tanulsingh077/twitter-sentiment-extaction-analysis-eda-and-model> - sentiment analysis python
* <https://learn.microsoft.com/en-us/power-bi/> - Power BI dashboard
* <https://developer.twitter.com/en/docs/twitter-api> - Data pull from Twitter
* <https://learning.postman.com/docs/getting-started/introduction/> - Checking an API call data in Postman
* <https://www.analyticsvidhya.com/blog/2022/01/text-cleaning-methods-in-nlp/> - understanding of text cleaning methods
* <https://www.youtube.com/watch?v=FVVgQdznjG0> – to understand visualizations that can be used for player analysis dashboard
* <https://dribbble.com/shots/6336391-Football-player-statistics-dashboard> - For dashboard designs
* <https://realpython.com/nltk-nlp-python/> - NLTK package used for text mining
* <https://www.researchgate.net/publication/284887142_An_Analysis_of_Best_Player_Selection_Key_Performance_Indicator_The_Case_of_Indian_Premier_League_IPL> - Paper that shows the methodology to select an ideal player for IPL (Indian Premier League)
* <https://www.sciencedirect.com/science/article/pii/S2772662222000029> - analytics used for player evaluation in T20 Cricket

## **Additional Data and analysis**

Following are the improvements that can be made to the current framework to make it more useful for real world applications

1. Include datasets from other social media networks, such as Facebook, Instagram image data, etc., to the existing data. More stats data from other sports websites can be added to improve the analysis and make them more accurate
2. Incorporation of Machine Learning models like LDA for topic modelling and other text matching algorithms. This will aid in providing more directional insights and recommendations
3. Automating the process. Currently the datasets are pulled for each date and the data wrangling process is done for a particular dataset. It would be ideal if all these processes were automated using ETL techniques along with integrity checks at each check point, so that the data is refreshed regularly with no errors and the production dashboard can be published with high quality
4. Analyzing and coming up with more creative graphs which provides insights across multiple angles