Project Plan

Sentiment Analysis Dashboard for Football and Cricket World Cup - 2022

**GROUP NUMBER 10**

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

**TABLE OF CONTENTS**

[Topic and Business Objective 2](#_Toc117108306)

[Information quality 3](#_Toc117108309)

[Data from API calls can have the following data quality/ Information quality issues 3](#_Toc117108310)

[Methods and tools 4](#_Toc117108311)

[External supporting materials 5](#_Toc117108312)

[Potential analysis and results 6](#_Toc117108313)

[References 6](#_Toc117108322)

## **Topic and Business Objective**

## Sports has always been one of the most spoken about topics in most of the social media platforms and there are plenty of people who ardently follow sports and want to understand deep nuances and sentiments for different kind of games and players. Our project will focus on developing an interactive dashboard that serves as a one stop solution for all sports enthusiasts to get a basic understanding about the popular posts and tweets about the upcoming Football and Cricket T20 word cups. This dashboard will help audience to follow/bet on the right kind of player for a particular match, help sports enthusiasts to gain a fair knowledge on most spoken about topics and concepts around the world cup matches, Sentiment around different players and teams across different geographies etc. It will be a fun dashboard for sports followers to get a quick overview of the latest sports updates.

## **Data** Social media Data is utilized for this analysis and will be used to understand the sentiment for different sports teams and players for cricket and football across different attributes like locations, audience types etc. for the upcoming World Cup. The data is pulled from Twitter and Reddit via API calls.

**Twitter Data**

Twitter Data is pulled using API call which requires API access keys

* API key
* API Secret Key
* Bearer Token

Using these creds, the Twitter API is hit, and the relevant data is pulled based on topics, subtopics, date interval and hashtags.

**Reddit Data**

Reddit Data is pulled using Reddit API and it doesn’t require API credentials for access to data.

The columns required for the analysis can be chosen from the API documentation provided.

* **Twitter API documentation -** [**https://developer.twitter.com/en/docs/twitter-api**](https://developer.twitter.com/en/docs/twitter-api)
* **Reddit API documentation -** [**https://www.reddit.com/dev/api/**](https://www.reddit.com/dev/api/)

Attached below is the sample of Twitter Data ***(Condition: all tweets containing trump in the recent past)*** and sample Reddit Data ***(Condition: Topic Politics)***

[Data Google Drive Link](https://drive.google.com/drive/folders/1wy52w40VeU5nN94NXdPbevoSMhRKgCa_?usp=sharing)

## **Information quality**

## **Data from API calls can have the following data quality/ Information quality issues**

* Every API call will have a maximum set of records it can pull per API hit and other time limits
  + Twitter API can be called only 15 minutes once
  + Reddit’s PushShift API only pulls 100 records per hit
* Twitter API cannot be accessed without API credentials and Secret Keys and there are different kind of accesses with each kind of access having separate privileges
* The API documentation and access formats keep changing and need to constantly keep checking the structure of API calls
* The data from API calls are unstructured. It is pulled as a **JSON object** which needs to be formatted and structured to give it a schema and **convert it as a dataframe object**

Graphical user interface, text, application, email

Description automatically generated

In the above image, the request is to pull all data having trump in the tweet and the result is highlighted in the rectangular box. This kind of data needs to be formatted and converted into Dataframe object

* There will be a lot of missing data in different fields and a lot of deleted comments and tweets

## **Methods and tools**

Graphical user interface, website

Description automatically generated

The Analytical Approach consists of the following steps

1. **Data Collection and Preparation** – The relevant sports related data is collected from Reddit and Twitter via API calls. Tweets and posts containing topics related to sports (Cricket and Football), World Cup and player related posts will be filtered and used for the analysis
2. **Exploratory Data Analysis** - After pulling the required data for a particular date frame, the data is explored for different patterns and trends, assessment of columns and subsets of data required would be deduced
3. **Data Wrangling Process** 
   1. **Data formatting** – Providing structure to the data, converting different Json objects to DataFrame objects, removing unnecessary attributes/columns
   2. **Missing Value Treatment** – Many columns will have missing values and these records need to be removed or replaced with a suitable value
   3. **Outlier Treatment –** Values that seem extreme and irrelevant to the analysis will be removed or identified in this process
   4. **Merging and combining datasets –** Both Reddit and Twitter datasets can be merged/combined based on any common columns like (Date, Location etc.)
   5. **Feature engineering**
4. **Potential Analysis**

Sentiment Analysis– Identify different sentiments for different players and Teams across different available attributes like location, age groups, gender etc. Identify popular opinions and popular tweets and comments for the sports and world cup topics

1. **Outcome – Dashboard** – The results from sentiment analysis will be showcased in the form of interactive graphs and visualizations in a Power BI

**Tools and Software:** The tools and software that will be used are

* + Power BI – Visualizations
  + Excel and Python – Wrangling and API calls
  + Postman – API calls and checks

**Challenges for the analysis**

* Segregation of complex text data and using it for analysis is tricky
* Change in API documentation can be a pain point since the structure of API calls need to be redesigned and re-run

## **External supporting materials**

* **Twitter API documentation -** <https://developer.twitter.com/en/docs/twitter-api>
* **Reddit API documentation -** <https://www.reddit.com/dev/api/>
* **Predicting wins in Premier League using twitter posts -** <https://www.sciencedirect.com/science/article/pii/S0167923616300835>
* **Understanding sentiment of soccer fans in US during World Cup 2014 -** <https://www.sciencedirect.com/science/article/pii/S074756321500103X>

Based on the Twitter and Reddit API documentation, several API calls and Sample data (JSON objects) on different topics and time periods were pulled and converted in DataFrames for analysis. Few twitter snippets are as follows

* Query with all tweets having world cup in it (Recent data)

<https://drive.google.com/file/d/1HXbPABAFDlHPhoJnN_jETlE2gsQO96aT/view?usp=sharing>

* Query with all tweets containing the Cricket Player “Virat Kohli”

<https://drive.google.com/file/d/11wSL2pu6VjAYSxQ7k9BFxx9ws-_wedQS/view?usp=sharing>

## **Potential analysis and results**

## Sentiment Analysis on different World Cup Teams and Player related topics and the results are displayed in the form of visualizations in Power BI dashboards.

## **Graphs and visualizations**

## Player vs number of tweets and posts

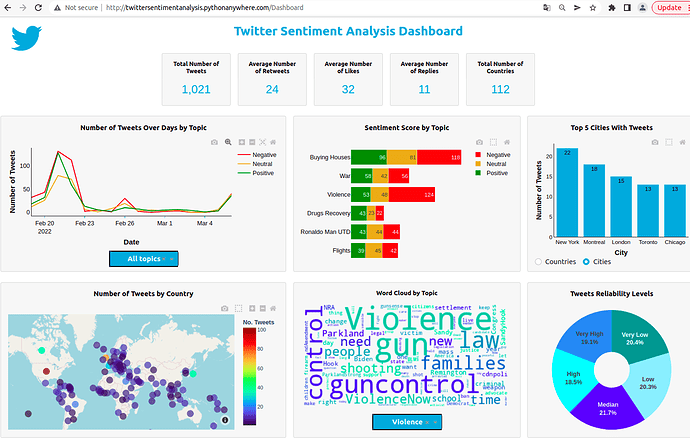
## Team vs number of tweets and posts

## Team with the maximum positive sentiment

## Team with the most negative sentiment

## Sentiment on Team and Players across locations, age groups, gender etc.

## Word cloud on the most spoken about topics and words



The interactive dashboard can be used to understand the sentiment across different players and Teams and can be leveraged by different **sports product companies** **–** to build products/merchandise with the trending quote and match wise news. Legal sports gambling companies and customers can use the dashboard to understand the player that has the most value and the team that has the most support.

## 

## **References**

<https://developer.twitter.com/en/docs/twitter-api/rate-limits> - Rate limits for Twitter