**BUSINESS DATA MANAGEMENT**

BY ALI KUMAYL

23110158

Contents

[**Introduction** 2](#_Toc121953332)

[**Data cleaning and extraction** 3](#_Toc121953333)

[**Sentiment Analysis** 8](#_Toc121953334)

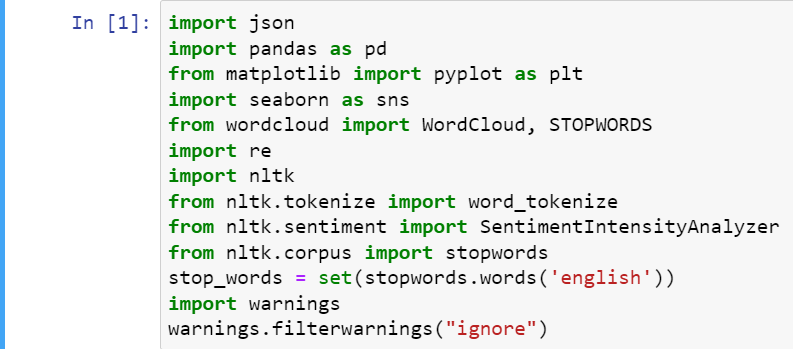
[**Conclusion** 13](#_Toc121953335)

### **Introduction**

A dataset was provided to us, with the file name ‘005.csv’ that contained information about different types of tweets. It contained tweets from 8th September 2022. We used Python and imported a number of libraries to help us with our data cleaning and data analysis. With the help of descriptive statistics and visualization, we were able to conclude that our dataset consisted of tweets mainly on the Asia Cup 2022 and Imran Khan. The dataset was unstructured so there were considerable inconsistencies that rendered it unreadable, difficult to understand, and difficult to interpret. This unstructured data set did not have any columns and the data was spread over numerous lines, rendering it hard to make sense of the data. The data was, therefore, cleaned up using different methods (as will be discussed later) and the end result was 100,000 rows and 20 columns. This means that our data consisted of 100,000 distinctive tweets along with 20 distinctive attributes which are:

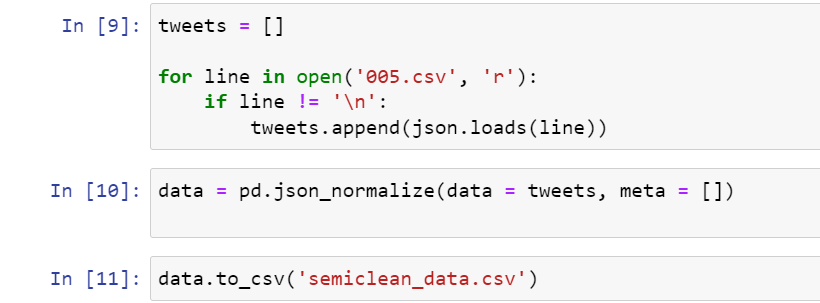
1. **Tweet time and date**
2. **Tweet ID**
3. **User ID**
4. **Protected**
5. **Verified**
6. **Followers**
7. **Friends**
8. **Favorites**
9. **Statuses**
10. **List Count**
11. **Account created at**
12. **User location**
13. **Quote count**
14. **Reply count**
15. **Retweet count**
16. **Favorite count**
17. **Text**
18. **Truncated**
19. **Extended tweet (full-text)**
20. **Language**

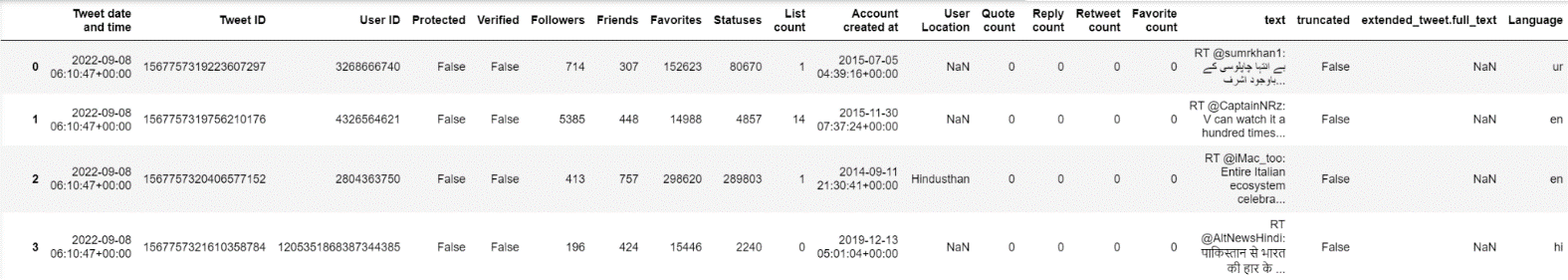
The libraries we imported for our data extraction and visualization are the following:



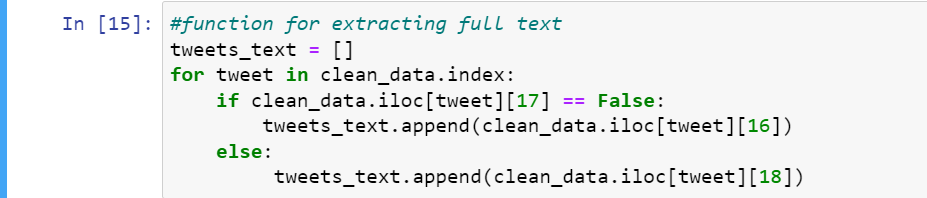
### **Data cleaning and extraction**

Since the dataset was widely jumbled and all over the place, it was essential to clean it using Python and extract data which was hidden within the file. The dataset provided to us was a json file, mainly unstructured data, therefore it was necessary to convert this data into a dataframe for analysis. Hence, we imported the json library in Python and used it to structure data as much as possible and then read it. For this purpose, we used the following code:

****

This gave us a data frame of 386 columns out of which we extracted and renamed 20 columns. Our final data frame looked like this:

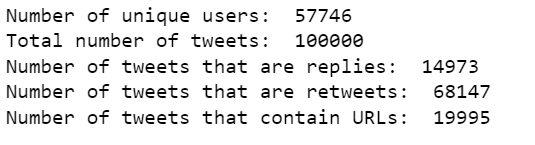
Another issue was that the text in the text column was truncated for longer tweets, meaning that if the tweet exceeded the limit, the entire full tweet was stored in another column. To extract the full text, we used 3 columns, ‘text’, ‘truncated', and ‘extended\_tweet.full\_text’. Code for this is shown below.

****

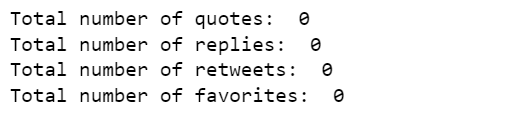
**Data Analysis and Visualization**

Visualizations are a good way to go when analyzing relationships in the dataset. We used the package Matplotlib and Seaborn in Python to assist us with this task.

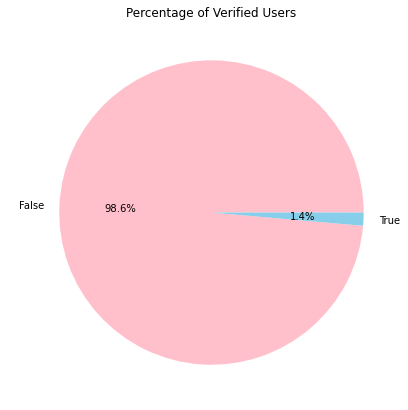
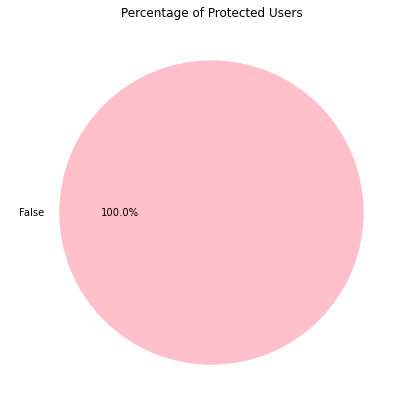
After structuring and cleaning our data, we used the help of descriptive statistics to understand the basic information about the users and their tweets. The figure below shows the overall statistics.

****

We also decided to check the number of tweets which had replies or were retweeted from this set of tweets. However, we were surprised to find out that none of these tweets were quoted, retweeted, replied to or liked. A sum of the corresponding columns showed zero for all columns, which is shown below:



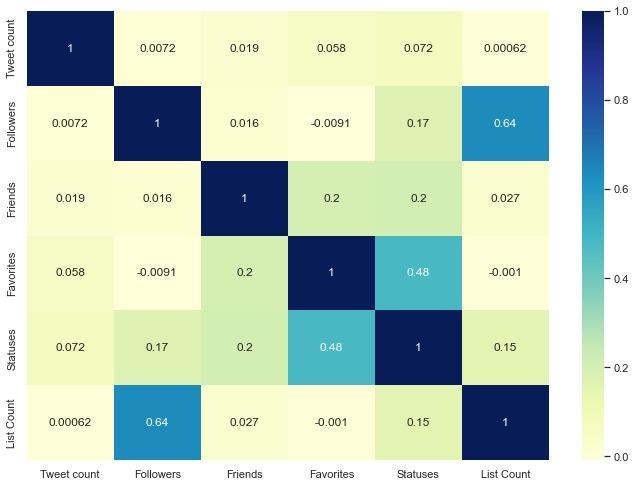
**Percentage of Protected and Verified Users**

****

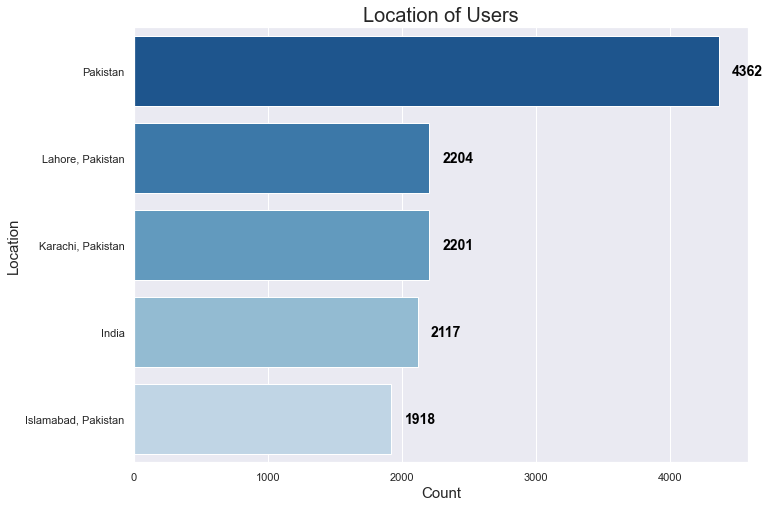
The percentage of protected users in the above pie chart is 100% false; meaning that all the accounts were public. The percentage ought to be 100% otherwise the tweet scraping wouldn’t have been possible if the accounts were private.

The pie chart for number of verified users shows that out of all 5776 unique users tweeting on 8th September, only 1.4% were verified users while a large percentage of 98.6% were unverified users. The gap, although huge, is not surprising. In fact, this tells us that the majority of the tweets were from the general population excited about the match or passionate about politics of Pakistan.

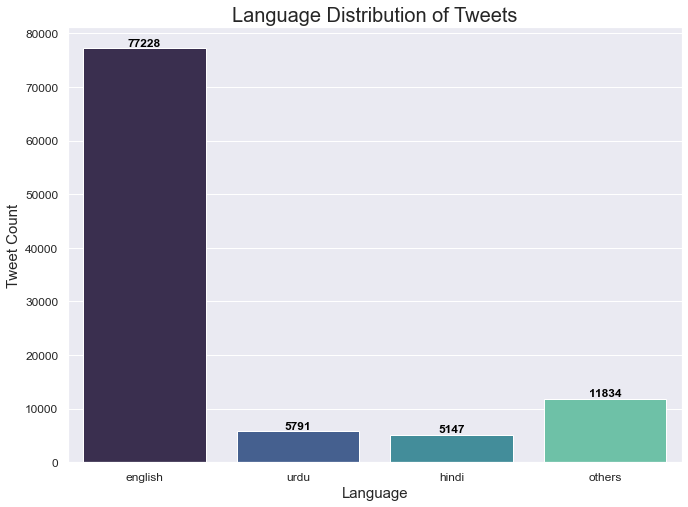
**Heatmap - Correlation Between User Statistics**

****This heat map indicates a high correlation between followers and list count with the correlation coefficient of 0.64. This indicates a general trend - greater the number of followers of a user, greater the number of public lists they are a part of. Another significantly positive correlation exists between favorites and statuses, with a correlation coefficient of 0.48. This was expected because higher number of statuses/tweets will generally mean that the cumulative number of likes on tweets (favorites) is higher. Apart from these, we expected that the correlation between favorites and friends would be high as well but the heatmap shows low correlation (0.2).

**Location of Users and Tweet Count**

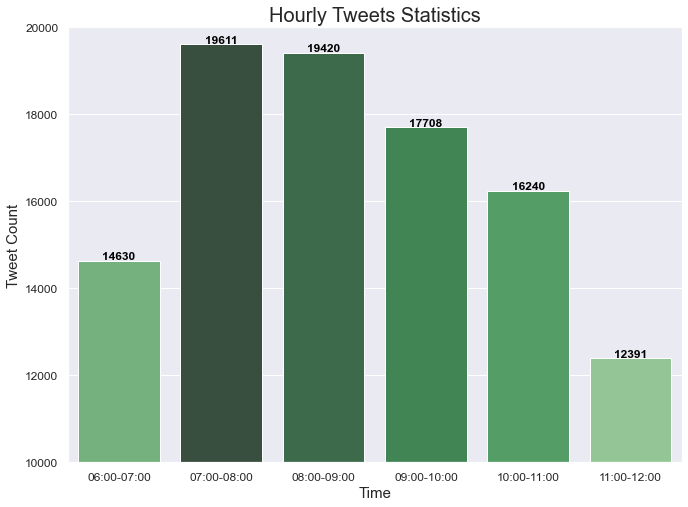
This graph shows tweets from the top 5 locations. Overall, 4362 tweets are from Pakistan, 2204 are from Lahore, 2201 from Karachi, and 1918 from Islamabad. The graph also shows significant tweets from users based in India. These figures correspond to the topic of discussion being the Asia Cup.

**Language and Tweet Count**

****

This bar chart shows the language distribution with relation to the number of tweets. When we used value.counts() function to count unique languages of tweets, we were surprised to see that there were 59 distinct languages used. English, urdu, and hindi were the most frequent, hence we decided to group all the others together for better visual representation. The results can be seen in the bar chart above. We can easily interpret this bar chart such that the users tweeted the most number of tweets in English, followed by others, Urdu and Hindi.

**Hourly Tweets and Tweet Count**

****

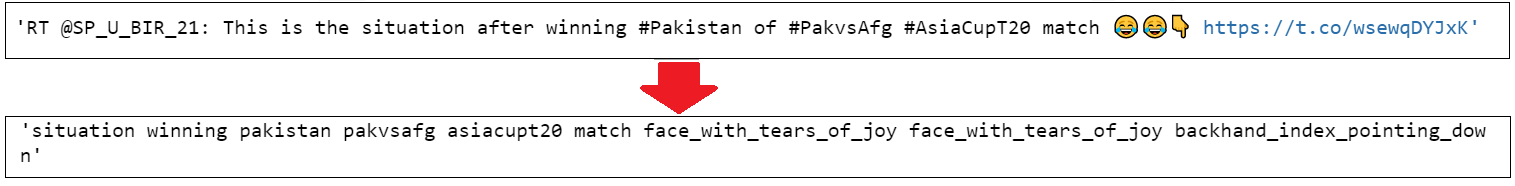
This bar chart shows the spread of the number of tweets over time. It can be interpreted as such that the number of tweets were highest between 7:00am and 9:00am (GMT) which can be translated to 12:00pm to 2:00pm PKT, indicating that users are most active during this time of the day. The lowest count is between 11:00am and 12:00pm (GMT) which is 4:00pm-5:00pm PKT, once again an indicator of user’s activeness on Twitter.

### **Sentiment Analysis**

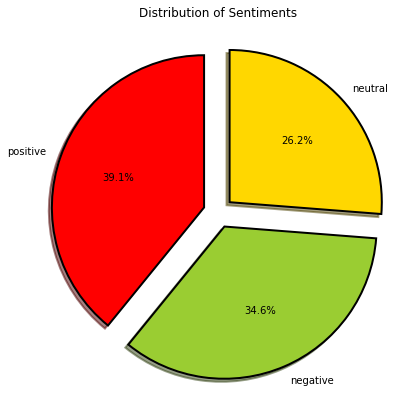
One of the limitations we faced while working on this dataset was associated with the use of different languages in tweets such as Urdu, Hindi and Arabic etc. We selected tweets in English language only for our sentiment analysis. After filtering english tweets, we created a function for cleaning tweets which is shown below.



This function removed URLS, hyperlinks, @mentions, hashtag symbol and RT (for retweets) and converted all text to lowercase. Emojis were converted into text too and then finally stop words were removed which gave us the final text for sentiment analysis. An example of unclean versus clean text is shown below:

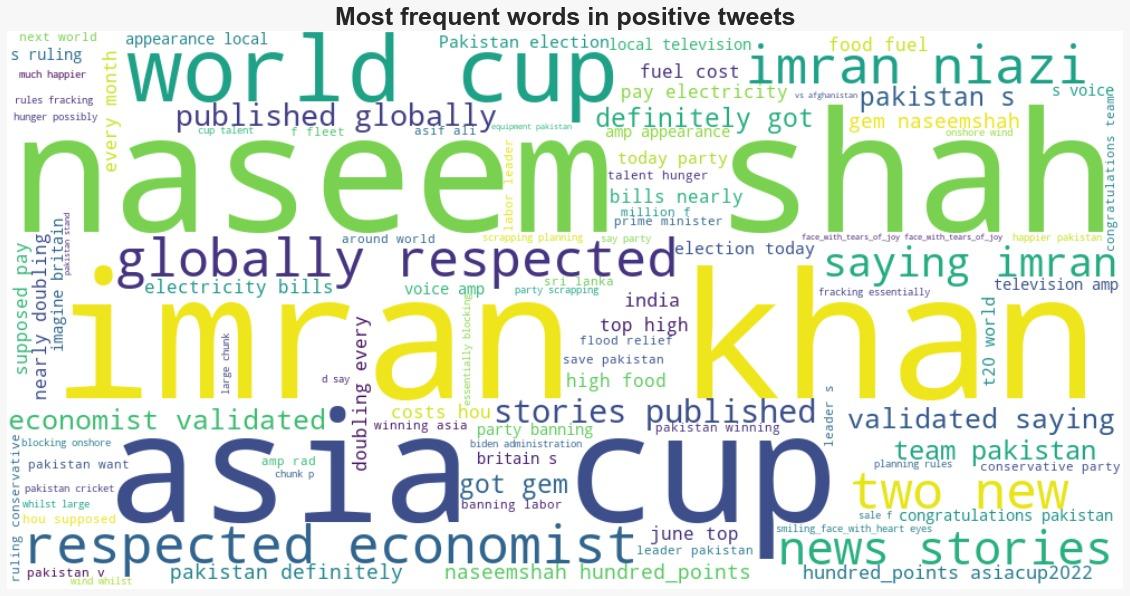


For each tweet, compound sentiment score was calculated, and the tweet was classified as either positive, neutral, or negative based on this score. The following pie chart shows the percentage of the three sentiments mentioned earlier.



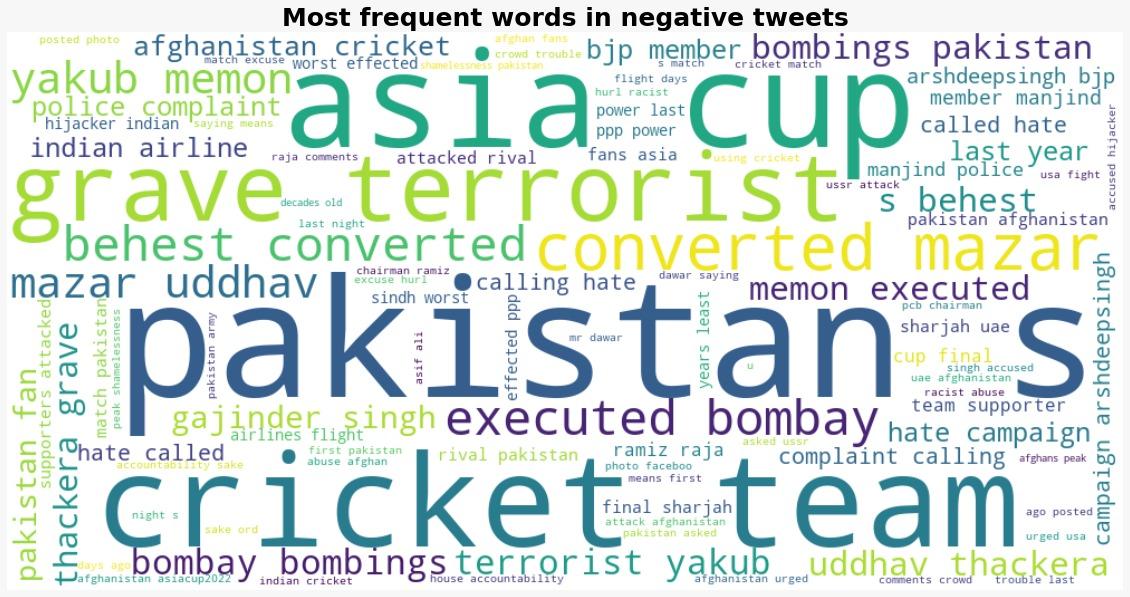
Tweets with positive sentiments accounted for almost 40% of the total tweets, while 34.6% and 26.2% were for negative and neutral sentiments respectively.

To further necessitate the understanding of the sentiment analysis, we used the help of visuals i.e. word cloud. When a word cloud is displayed, the most frequently used words appear larger or bolder than those around them, indicating the relative popularity of the word. The following word cloud shows the most frequent words used in positive tweets.



It is clear at first glance that the tweets were surrounding the Asia Cup in general and the Pakistan vs Afghanistan tournament in particular, which was held on 7th of September (words like Asia Cup, Naseem Shah and team Pakistan were used frequently). Apart from cricket being the topic of discussion, the word cloud shows positive words associated with the politics of Pakistan or to be specific Imran Khan.

The word cloud for the most frequent words used in negative tweets is displayed below:

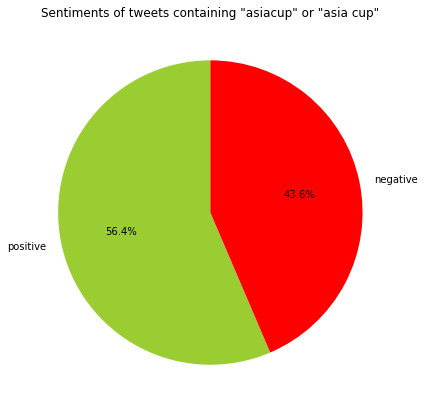


Surprisingly, the negative tweets also consisted of the same words used in the positive tweets i.e. Pakistan and Asia Cup. However, the word ‘bombay bombings’ and the events associated with it (words like terrorist, executed bombay and terrorist yakub) are spread out in the image, indicating that this 1993 incident was one of the most discussed topics on Twitter on 8th September.

Apart from this, the pie charts shown below also highlight people’s positive and negative sentiments associated with tweets which had the words ‘Pakistan’ and ‘Imran Khan’ in them.

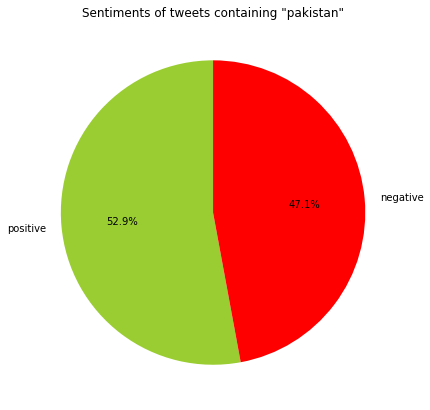
We decided to check sentiments around a few words – mainly those appearing in both negative and positive tweets.

**Sentiment around Asia Cup**



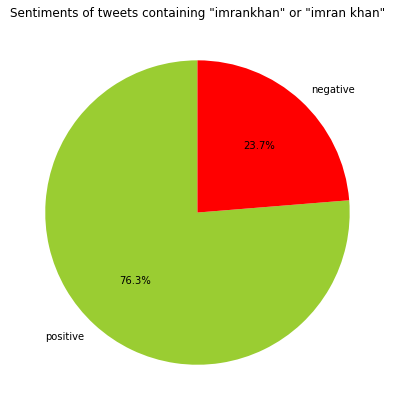
The Asia Cup of 2022 was the talk of the town. People all around Asia showed enthusiasm for the cricket matches and supported their teams through words of encouragement on Twitter as can be proven by the percentage of positive sentiments (56.4%). However, there were some negative sentiments found in 43.6% of the tweets probably owing to the fact that users tweeted against the opponents of their favorite teams using words and emojis with negative sentiments attached to them.

**Sentiment around Pakistan**

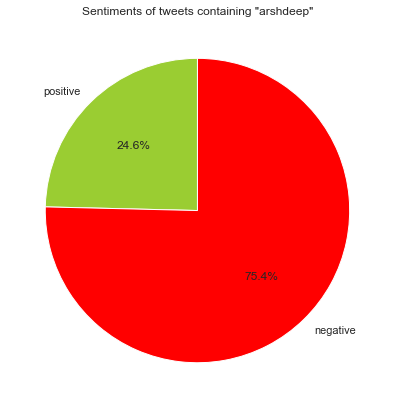


Tweets which included ‘Pakistan’ in them were 52.9% positive while 47.1% negative. The positive sentiments can be correlated to Pakistan’s performance in the cricket match with Afghanistan. Pakistan won the match and people’s positive sentiments for the Pakistan cricket team emerged in the tweets about them. Similarly, 47.1% of the negative tweets are in correspondence with the sentiments of non-Pakistanis (possibly Indians in this situation) who were not happy with Pakistan winning the match. However, one can claim that these sentiments can be in correspondence with the politics of Pakistan, this may be true, but there is no way to distinguish between the two topics of discussion.

**Sentiment around Imran Khan**

To talk about politics as being the topic of discussion in the majority of the tweets, we have used the above pie chart to categorize the sentiments of people surrounding Imran Khan. This pie chart shows 76.3% of the tweets with ‘Imran Khan’ in them were positive while 23.7% were negative. We can correlate these sentiments with the Jalsa of Imran Khan held on 7th September in Chistian which had some political significance. These insights are an interesting way for politicians to gauge their popularity using Twitter’s API services, as can be proven by these analyses.

**Sentiment around Arshdeep Singh**

An interesting finding during our analysis was that people continue to tweet about a significant event for several days. The pie chart above shows sentiments of tweets containing the name ‘Arshdeep’. 75.4% of the tweets were negative in nature for this Indian cricket player. This can be linked to India’s defeat in the cricket match against Pakistan on 4th September, where Arshdeep dropped a very important catch of Asif Ali, turning the game around.

### 

### **Conclusion**

Twitter is a great source for data collection as the users mostly have public accounts with their personal information accessible to all. Twitter API lets users read and write data, and it can be used to get followers' data, compose tweets, and read profiles as well as access a large quantity of tweets on certain topics in certain locations. Users can use Twitter API to gather data on a particular subject. In fact, a lot of companies use Twitter’s API services to access real time data and make inferences using historical data as well. Once the data is obtained, it can easily be interpreted using Python and its tools.