PSL TWEETS

An Analysis

Group 04

MALIK MUSSAB	23110229
NAMRA ASGHAR	23110034
RUBAB MUKHTAR	23110226
FARAN MAOOD	23110100

TABLE OF CONTENTS

ABSTRACT

“

Using the Twitter dataset provided, the project goes over the process of cleaning the data to convert it into a usable form through which further analysis can take place. The project further focuses on identifying trends and sentiments of the public that can be deduced through their tweets during the season of PSL.

”

INTRODUCTION

For this project, we were given twitter metadata for daily tweets targeting the keyword “**PSL**”. The dataset contained tweets from ***Feb 27, 2022***, timed from ***17:15:54*** to ***18:06:40***. According to the schedules on the PCB website, this was the day of the Final Match of PSL 2022: Lahore Qalandars vs Multan Sultans. The inclusion of multiple languages including roman language (Urdu written in English) in this dataset made this project unique as conventional methods alone were not enough for interpretation and further analysis and assumptions were required for us to understand the trends in data. We went through a journey of data cleaning followed by an assessment of what areas to specifically study for our project after which we were able to deduce trends and patterns and understand the public sentiment regarding PSL.

METHODOLOGY

We took the approach of exploratory research in this project as we were experimenting with a type of dataset that was different from the usually undertaken projects. Throughout our analysis, we found that we mostly conducted Qualitative research as the sentiments and emotions of the audience were being analysed. However, the quantitative part of the project was of great importance as well as that provided a strong foundation for our qualitative analysis since our numbers went in line with the trends that were inferred.

DATA CLEANING

Data cleaning is a crucial step in the data science pipeline as the insights and results one produces is only as good as the data you have. As the adage goes — garbage in, garbage out. Using unclean data to produce analysis will result in erroneous predictions that engender bad decisions and dangerous outcomes. Most machine learning algorithms only work when your data is properly cleaned and fit for modelling.

Steps Taken

- **Loading raw file:**

The first step inculcates loading the file from the directory to the interface.

```
import pandas as pd
df_1 = pd.read_csv("Group_4.csv", usecols=[0], names=['created_at'])
```

- **Extracting attributes:**

After loading the dataset, extraction with regards to attributes is begun:

1) Tweet Date

clean_date function was defined to extract the Tweet Date attribute using the **re.sub**, created as an alias for replacing the garbage data (symbols primarily) with blank spaces. The date was converted from string to **date time format** through the **datetime library**.

```
def clean_date(dates):
    str(dates)
    dates = re.sub('{"created_at":', '', dates)
    dates = re.sub(r'\+0000', '', dates)
    dates = re.sub('Sun', '', dates)
    dates = re.sub('"', '', dates)
    return dates
```

2) Tweet ID

The IDs in the data were starting with "id: ", hence the **clean_id function** was created and used to clean the prefix to just an id number in the **integer** format.

```
def clean_id(id):
    id = re.sub('id:', '', id)
    return id
```

3) Sources

To extract the sources used for tweeting, an **empty list** was defined and created. A **for loop** was run across the data which iterated over rows and extracted the five sources via the employment of **if conditionals** and **str.find()** to find which source was used and **while** to extract the mention of the source from ':'.

4 - 18) All Other Attributes

15 attributes were **extracted together** because their data was not distinctly separate in different columns, rather it was all merged with each other. The columns in which all this data was contained were defined as a **range** from 11th to 233rd index. Using **df.iterrows()** function, all the rows were iterated upon and **split()** function was used to extract useful data. All these were added to separate distinct columns.

```
for i, row in df.iterrows():
    j = 0
    while(row.values[j].split(":")[0] != "user"):
        j += 1
    user_id.append(row.values[j].split(":")[-1])
```

- **Cleaning Tweet Text:**

Urdu Language

Tweet text included phrases and tweets in Urdu and Roman Urdu which made it unusable for sentiment analysis. Thus, a function with the name **remove_lang** was created to make use of the **len() function** and remove phrases which began with **\u** till the end and replace them with a blank space.

```
def remove_lang(tweet):
    new_tweet = []
    for word in tweet.split():
        allowed = True
        for i in range(len(word)):
            if(word[i:i+2] == r'\u'):
                allowed = False
        if allowed:
            new_tweet.append(word)

    new_tweet = " ".join(new_tweet)

    return new_tweet
```

Extra Symbols

These were removed in order to make the text data useful. For this, the **clean_tweet** function was created and was used with **re.sub()** which catered to symbols including *@, http\, hashtags, retweets, stop words, punctuations, URLs, texts* in languages other than English and mentions. However, the initial data made use of @ to gauge replies and URLs for initial data analysis purposes. These were removed for sentiment analysis.

Functions used to do these included **remove_punct**, **remove_stopwords** and **string_punctuation** which has a list of all punctuations and replaces them with blank spaces.

DATA LOADING

After extracting variables and important columns, the csv file was loaded to the **jupyter** notebook in a **data frame** using the **Pandas library**. It was used to conduct data exploration due to the multitudes of built-in functions it has that add to its usability.

```
#!/usr/bin/env python
# coding: utf-8
import pandas as pd
data = pd.read_csv("Final_Cleaned_Data.csv")
data.head()
```

It helped in running queries to relate variables, find general trends and distinguish the distributions of variables to make sense of the data. Visualisations were incorporated via the use of Microsoft **Power BI** discussed later. Each query has been explained with the output shown.

Attributes

Total number of attributes extracted = **18**

1. **Tweet_id**
   Unique ID of the tweet
2. **Tweet_date**
   Date and time when the tweet was posted
3. **Source**
   Twitter platform used by users to post the tweet
4. **Quote_Count**
   Number of quote tweets on the user's tweet
5. **Reply_Count**
   Number of reply tweets on the user's tweet
6. **Retweet_Count**
   Number of times the user's tweet was retweeted
7. **Tweet_Favourite_Count**
   Number of times the user's tweet was liked
8. **Tweet**
   Text of the tweet itself
9. **User_id**
   Unique ID of the creator of the tweet
10. **Name**
    Name of the creator of the tweet
11. **Protected**
    Whether the account of the user is protected or public
12. **Verified**
    Whether the account of the user is verified
13. **Followers_Count**
    Number of users following this Twitter user
14. **Friends_Count**
    Number of other Twitter users this user is following
15. **Favourites_Count**
    No. of tweets user has liked since the account's creation
16. **Statuses_Count**
    No. of tweets this user has made since the account's creation
17. **Created_At**
    Date & time of account's creation
18. **Location**
    User's reported location where the tweet is posted from

DATA EXPLORATION

Total Number of tweets

20,000

- Since all the rows were unique with their own Tweet IDs, the **row count** was found using the **len()** function to calculate the total number of tweets.

Timeframe of the tweets

Date = **27/02/22**
Timing = **17:15 to 18:06 hours**

- The dataset was sorted on the variable Tweet_date in an ascending order to obtain the first and the last time the tweets were posted.

- The results show that the entire dataset is based on one day within a span of **52 minutes**.

Number of Retweets

304,054

- The variable Retweet_Count, extracted in the data cleaning step, gave the number of times a tweet was retweeted. It was summed to calculate the total number of retweets.

Number of Unique Users

9,154

- The variable User_id, extracted in the data cleaning step, gave the unique id for each user. The function **nunique()** was applied to it to retrieve the number of distinct elements in the variable.

URL-containing tweets

7,020

- Since URLs begin with '**http**', the Tweet texts were converted to strings and the occurrences of 'http' in the tweets were counted. The counts were summed up and converted to an integer value.

Number of tweets that are replies

10,112

Average Number of Characters and Words per tweet

Characters = **57.43**

Words = **8.24**

- To calculate the average number of characters per tweet, the len() function is used to find the length of each tweet - which basically gives the total number of characters, and the lengths are then averaged out using the **mean()** function.

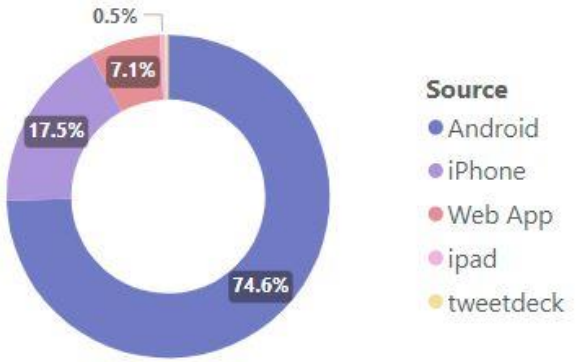
- To calculate the average number of words per tweet, first, the **str.split()** function is used to split the tweet string into a list containing words as items. Next, the length of the list is found by the len() function for each tweet - giving the total number of words per tweet. The totals are averaged out using mean().

- The means are rounded up to 2 decimal places by **round()** function to yield clean figures.

Twitter Platforms used to tweet

1. Android = **14,922**
2. iPhone = **3,491**
3. Web App = **1,425**
4. iPad = **97**
5. Tweetdeck = **65**

- The **value_counts()** function is applied to the Source attribute to find the total number of times each platform was used.



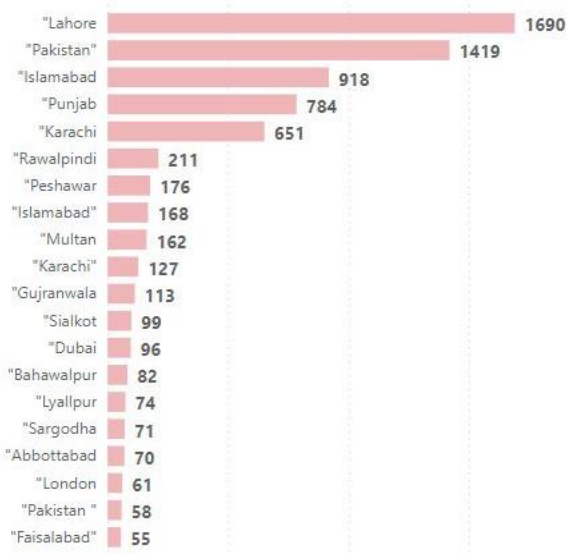
User-Reported Locations of tweets

- The **top 20 locations** of the tweets in the dataset are as shown in the graph below:

- There are a lot of **overlapping** and **repetitive** locations in this variable: Although most of the elements are cities like “Lahore”, “Islamabad”, and “London”, there are a few countries like “Pakistan” and “England”, and provinces like “Punjab” as well.

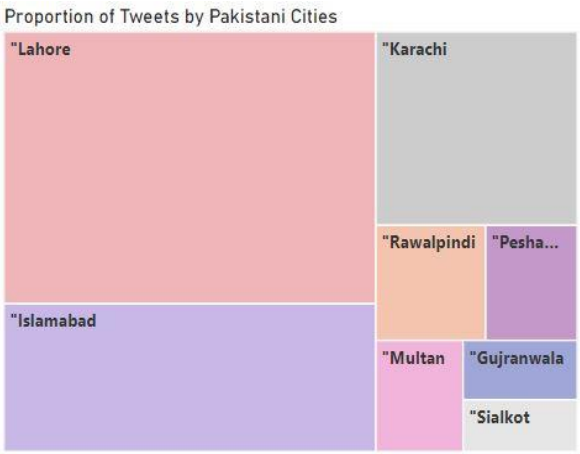
- Hence, the numbers depict an **inaccurate variation** in the locations of the users.

- The **value_counts()** function sorted in descending order is applied on the Location attribute to find these.

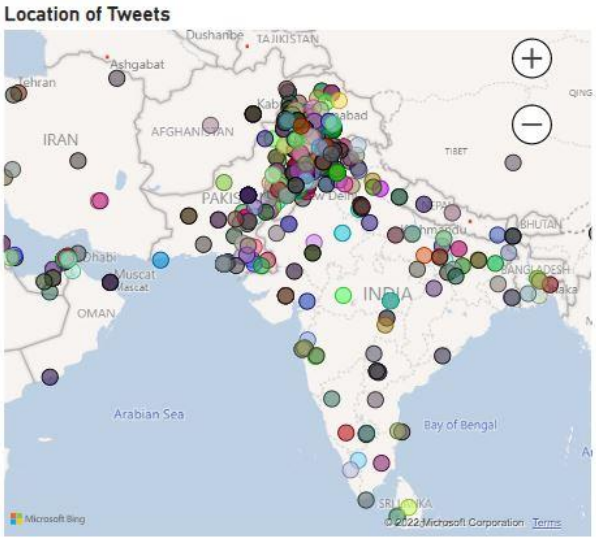


Top Pakistani User-Reported Cities

1. **Lahore** = 1,668
2. **Islamabad** = 910
3. **Karachi** = 644
4. Rawalpindi = 210
5. Peshawar = 176
6. Multan = 162
7. Gujranwala = 113
8. Sialkot = 99
9. Bahawalpur = 80
10. Lyallpur = 74
11. Sargodha = 71
12. Abbottabad = 70
13. Faisalabad = 53
14. Okara = 45
15. Hyderabad = 38



- Although the **majority** of the tweets report a location in **Pakistan**, there were **some international** tweeters as well, including those in India as the map depicts.



International User-Reported Locations

1. **Dubai** = 96
2. **London** = 61
3. **England** = 46
4. United States = 25
5. India = 24
6. United Arab Emirates = 24
7. Abu Dhabi = 23
8. Kingdom of Saudi Arabia = 23
9. Kabul = 19
10. Toronto = 15

Highest and Lowest Number of tweets by a user

Highest = **73**
Lowest = **01**

- The value_counts() function sorted in descending order is applied to the 'User_id' attribute to find the highest. The **tail()** of the above data frame is viewed to obtain the lowest number.

Names of the Top 05 Tweeters

- The **groupby()** function is used to attach the 'Name' alongside the above data frame.

1. N/A = 73
2. Maica = 50
3. PSL Live Score! (BOT) = 40
4. Saqlain Maqsood = 38
5. Muhammad Umair Aslam = 34

Average Number of tweets per user

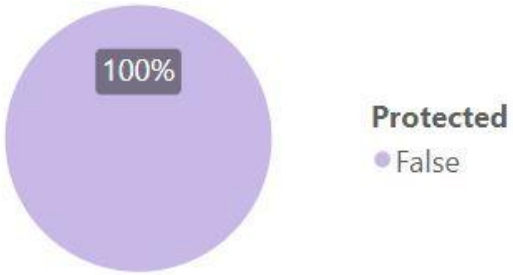
2.18 tweets

- The total number of tweets is divided by the unique number of users to find the average.

Number of Protected Accounts

True = 0
False = **20,000**

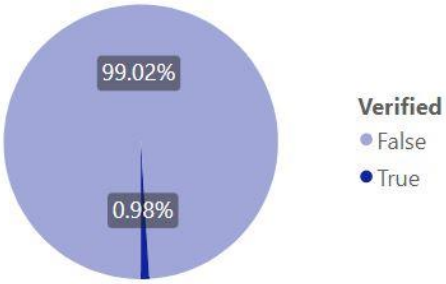
- All the tweets are from **public accounts** as no tweet holds true for protected accounts.



Number of Tweets from Verified Accounts

True = **196**
False = **19,804**

- **0.98%** of the tweets are from **verified** accounts
- **99.02%** from **unverified** accounts.

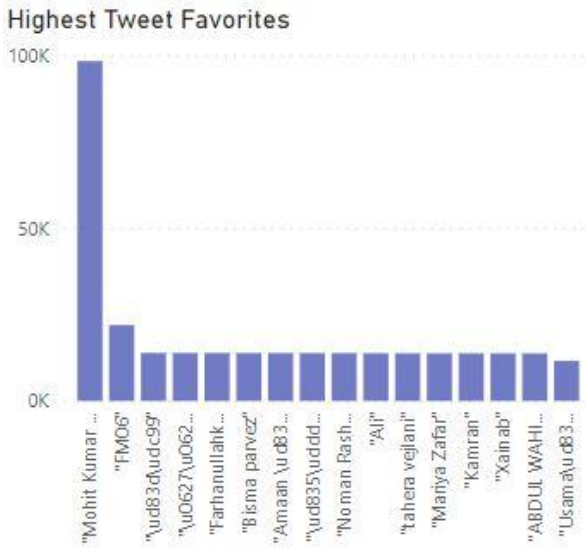


Tweet Favorite Count

Most Favorites on a Tweet

Name = **Mohit Kumar**
Favorites = **98,300**
Location = **Jaipur**

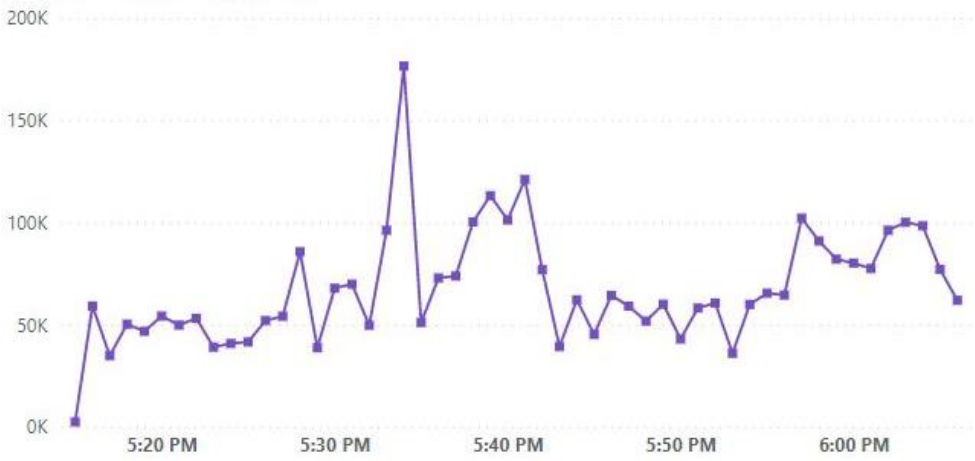
Tweet = ***"Look at the fear in those eyes"***



Timing of the Most Favorite Tweets

- The average timing at which the top 10 most favourite tweets were posted: **17:45**
- Of the 52 minutes of all the tweets, most favourite tweets fall in the **latter half** of the total timeframe. The average timing falls on the **30th minute**.

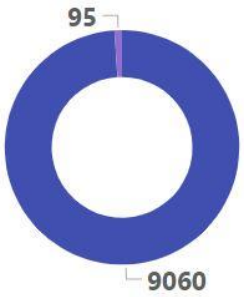
Trend of Favorites on the Tweets



Average Followers of Most Favorite Tweets' Tweeters

- The average followers of the tweeters of the **top 10** most favourite tweets = **770**
- The favorites are on average = **24,000** which is **31x times** the average number of followers. This means that these tweets reached 31 times more twitter audience on average than their usual followers.

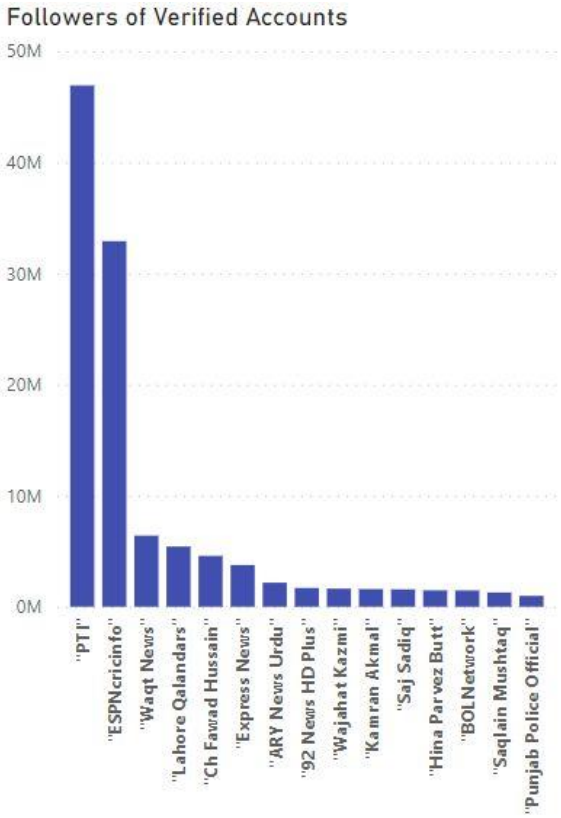
Verified Accounts



- Of the 9,154 unique users, **95** have **verified** accounts. = **1.04%**
- These 95 verified accounts belong to journalists, news channels, cricket reporters, celebrities, political parties and leaders.

Most Popular Verified Accounts

Based on the total number of followers of these 95 verified accounts, their popularity was determined. **Top 15** accounts are shown in the graph below.



- **PTI** is the **most popular** with **46.9m** followers.
- **ESPNcricinfo** comes **second** with 32.9m followers.
- The rest of the verified accounts have significantly fewer followers than the first two.
- Waqt News = 6.4 million followers
- Lahore Qalanders = 5.4 million
- Ch Fawad Hussain = 4.6 million
- Express News = 3.8 million
- Ranging from 1.3 to 2.2 million: ARY News Urdu, 92 News HD Plus, Wajahat Kazmi, Kamran Akmal, Saj Sadiq, Hina Parwez Butt, BOL Network, Saqlain Mushtaq
- Punjab Police Official = 980k followers

Looking at the Top 15 most popular verified accounts, it can be evidently seen that there are many **News Channels** which tweeted about this subject. Hence, an analysis of these accounts was conducted.

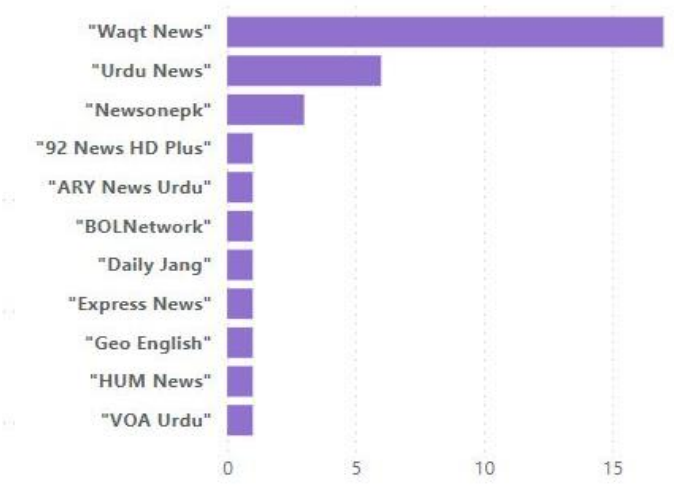
Analysing News Channels

By extracting names of the verified accounts containing the word “news”, “english”, “urdu”, and inspecting further, a **total of 11** News Channel accounts were identified:

1. Waqt News
2. Urdu News
3. Newsonepk
4. 92 News HD Plus
5. ARY News Urdu
6. BOLNetwork
7. Daily Jang
8. Express News
9. Geo English
10. HUM News
11. VOA Urdu

Number of Tweets by News Channels

- **Waqt News** has tweeted **17** times - the **highest**
- Urdu News which has tweeted 6 times
- Newsonepk has tweeted 3 times
- All other news channels have tweeted only once in the timeframe



Popularity of News Channels

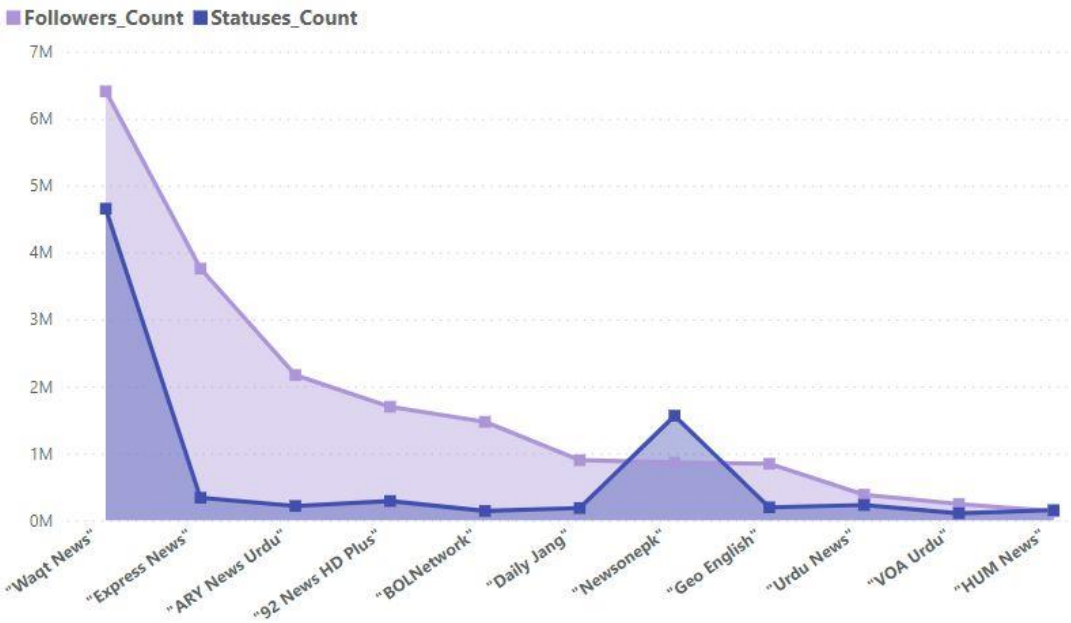
To assess the popularity of news channels, the total **number of followers** of the 11 News Channel accounts has been graphed. It can be seen that:

- **Waqt News** has the **highest** number of followers (**6.4 million**)
- Followed by Express News with 3.8 million and ARY News Urdu with 2.2m followers.
- All other news channels have under 2 million followers and some under 500k.

To determine whether the number of times a news channel tweets has any impact on the number of followers it has, the total **number of statuses** of all news channels was graphed:

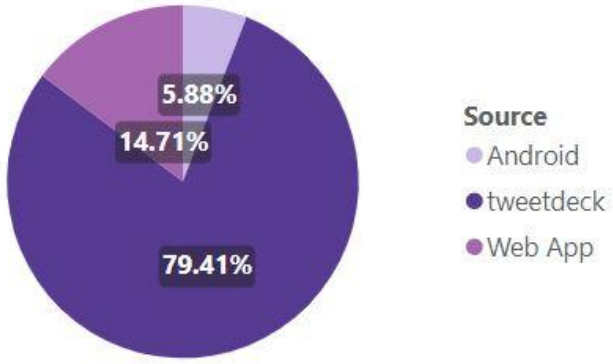
- The **trend** for the number of followers and the total number of statuses is almost **similar** as shown by the two graphs together below.

- A high number of statuses shared is **positively correlated** to having a high number of followers.



Twitter Platforms used by News Channels

- **TweetDeck** was used the **most** (**79.41%**) times to tweet out of the 5 sources (Android, iPhone, Web App, iPad, TweetDeck) available for tweeting.
- **Web App** was used 14.71% times.
- 5.88% of tweets by Android.
- No tweets by News Channels have been done by iPhone or iPad.

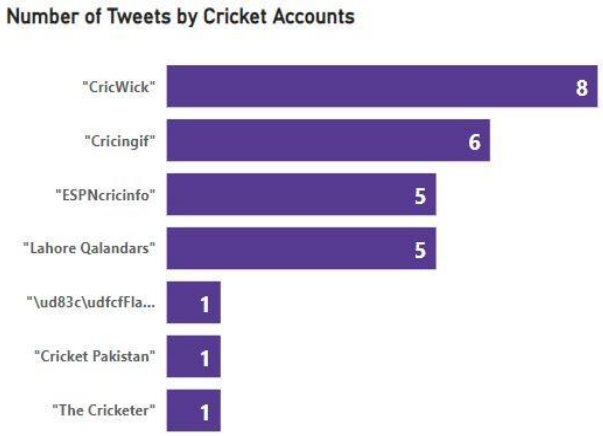


As the subject under discussion is Pakistan Super League (PSL) 2022, looking at the verified accounts, there are many **Cricket Accounts** which tweeted about it. Hence, an analysis of these accounts was conducted:

Analysing Cricket Accounts

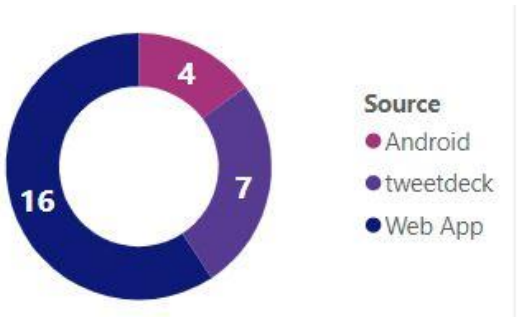
By extracting names of the verified accounts containing the word "cric", and inspecting further, a **total of 07** Cricket-related accounts were identified:

1. CricWick
2. Cricingif
3. ESPNcricinfo
4. Lahore Qalandars
5. Flashscore Cricket Commentators
6. Cricket Pakistan
7. The Cricketer



Twitter Platforms used by Cricket Accounts

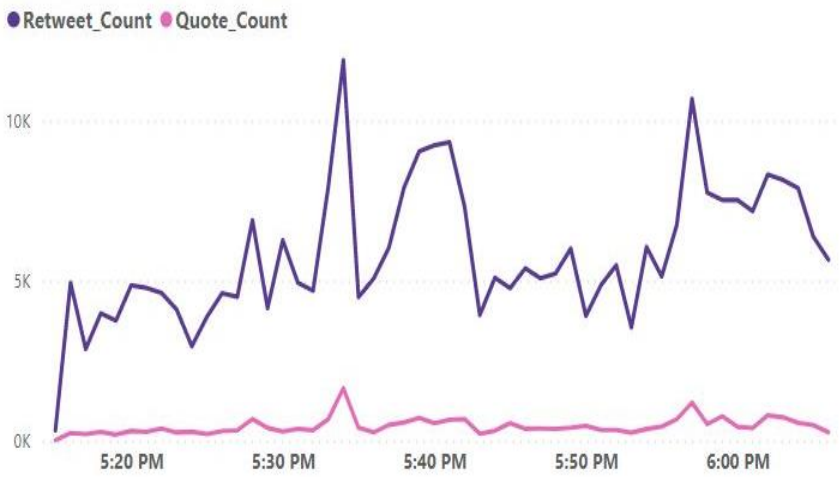
- **WebApp** was used to post **16** tweets by Cricket accounts.
- **TweetDeck** was used **07** times.
- **Android** was used **04** times.
- No tweets by Cricket Accounts have been done by iPhone or iPad, just like in the case of News Channels.



Trends Over Time

Retweets and Quote Tweets

- The graph shows the time leading up to a match which indicates the building up of **hype** before a match.

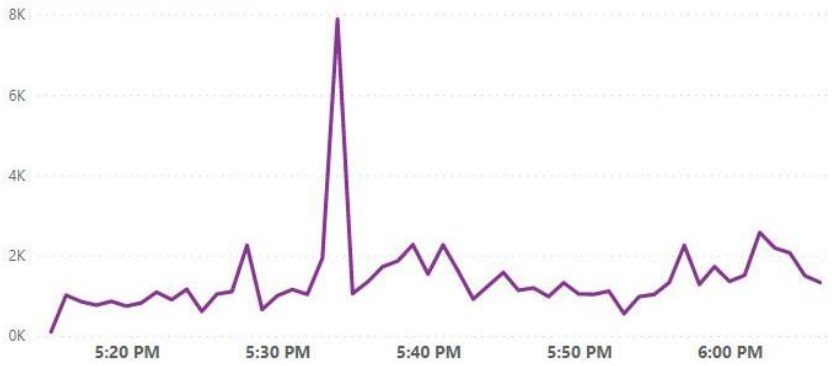


- This is further supported by the **greater number** of **retweet** counts rather than quote counts which further supports how fans share their supporting team's tweets out of **impulse**.

- Another interesting trend can be noticed in the **peak of both** Retweet count and Quote count between 5:30 PM and 5:40 PM which is reflective of the greater engagement of the audience during pre-match analysis by experts such as Shoaib Akhtar, Wasim Akram, and Waqar Younis.

Replies on Tweets

- The replies on tweets overall have a **comparatively lower** number compared to the retweets. These can mainly be attributed to Twitter's ease of use in resharing tweets compared to replying.

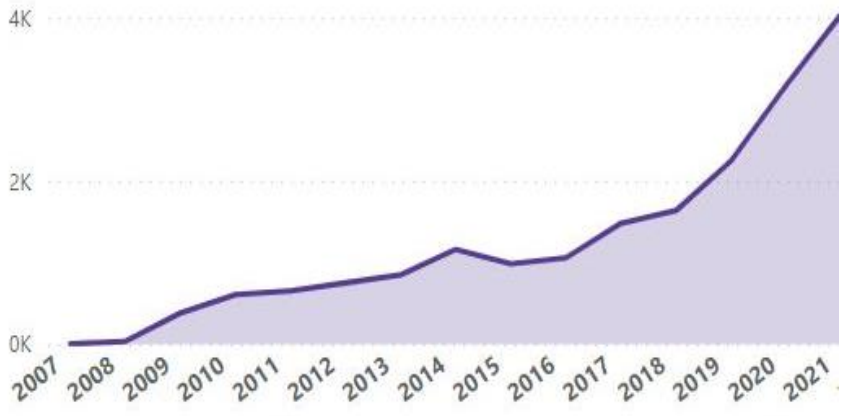


- Here, we can also notice the **peak** between 5:30 PM and 5:40 PM but compared to the retweets, the peak for replies is much **higher** when compared to the normal graph.

- This can again be associated with the expert's pre-match analysis, but which gains more traction in terms of people sharing their own opinions about the analysis regardless of whether they agree or disagree with the analysis.

Growth in Twitter Accounts

- Through this graph, we can depict that the tweeters on this subject belonged more to those who **recently created** their accounts. the growth of Twitter accounts has a direct effect on the **increase** in tweets, replies, and reshares.
- However, it must be noted that the increase in the number of accounts does not have a 1:1 effect on the tweets which is indicative of **dormant accounts**.



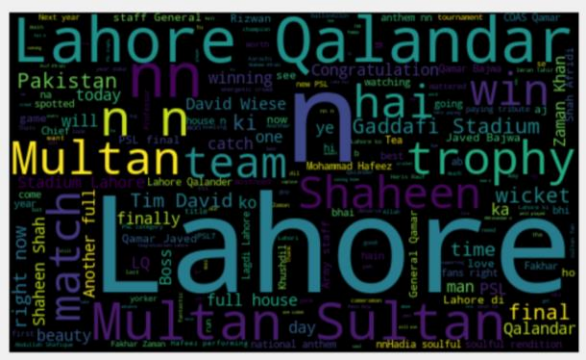
SENTIMENT ANALYSIS

- The method of 'computationally' assessing whether a piece of text is good, negative, or neutral is known as sentiment analysis. It's also known as opinion mining, and it involves determining a speaker's viewpoint or attitude.
- Sentiment analysis can be used to monitor and analyse social phenomena spotting potentially dangerous situations and determining the general mood of the blogosphere.
- In this dataset, we were provided with data pertaining to PSL, and the objective was to understand people's sentiment regarding the event on Twitter.
- The most important step for sentiment analysis is to make sure the text data inserted is clean and has minimal noise, so as part of the data cleaning steps, it was ensured that the unnecessary text was removed from the tweets, which included removing:
    - Hashtags
    - Retweets
    - Stopwords
    - Punctuations
    - Stopwords
    - URLs
    - Text in languages other than English
    - Mentions

- Following this TextCloud library was loaded, and functions were defined that iterated over the entire tweet text rows and returned the subjectivity of the tweet, which measures if the tweet is subjective or objective. In addition to this, the polarity of the tweets was also calculated, which measured if the tweets expressed negative or positive emotions.

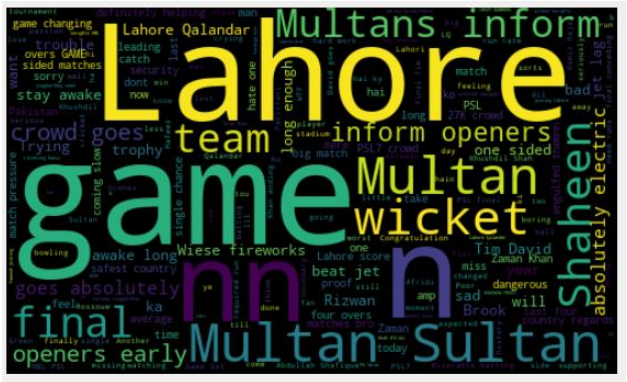
- Moving on, the WordCloud library was imported to generate word clouds of the data. The size of each word indicates its frequency in the dataset and hence its importance. So, the bolder and bigger a specific word appears in the word cloud, the more significant it is in the dataset. In the data set cleaning part, useless and meaningless words were removed so that they do not appear in the word cloud and provide meaningless information.

Wordcloud with Overall Subjectivity Incorporated



Wordcloud with Positive Tweets only Incorporated

Wordlcoud with Negative Tweets only Incorporated



LIMITATIONS

- **Inaccuracy of User Locations**
  It prevented us from doing a detailed location-based analysis of the tweets. This limitation is caused due to the **comedic intention** of certain users who express their current state of mind as their current location. An example of this is a user who identified their location as "***mazay main***" (translation: chilling) which is not a location, but rather, the mood that the person tweeted in.

  If we were to carry on with location-based analysis on this data, large anomalies were expected due to which we deemed it fit to not discuss these trends in a very detailed manner to prevent any misreporting.

- **Timeframe of the Dataset**
  As PSL matches on the mentioned date were played **after 7 PM**, the timing of the data till 18:04 **did not allow us** to notice the changing sentiments and emotions of the audience during **gameplay**.

  This caused a major setback as much of the detailed analysis could be based on this since the unexpected changes during the game result in significant changes in the audience's sentiments which could result in very interesting results. It would be one of our top suggestions for experts intending to pursue this project further to include the additional timeframe to better analyze the data.

- **Urdu Text**
  We were not able to go in-depth on the data found in the **Roman** Language (Urdu typed in English) as **tools** for this use were **not available** with high credibility. This forced us to limit our dataset to those tweets that were easily used in the sentiment analysis. Since the Pakistani audience generally tweets in Roman Urdu, it would have been very beneficial for overall trend analysis if tools were available that would run sentiment analysis on tweets in Roman Urdu as well.

CONCLUSION

PSL received an overwhelming response from the audience where we were able to get a dataset of 20,000 tweets in less than an hour's time. Through detailed analysis of this data, we were able to determine how these tweets were not limited to a single income segment or city, rather, tweets were posted from all over Pakistan and even from other countries. Furthermore, the rise in popularity of Twitter brings together people from all income segments and professions where celebrities and news channels also engage in these discussions. Overall, the majority of the discussions found were of a positive sentiment that indicates the importance of the event in our country.