

**MDA 720**

**Final Project**

**“Rafa Nadal Tweets for his 22<sup>nd</sup>  
Grand Slam Title”**

**Sergio Fernandez**

## Contents

Introduction:.....	3
Problem Analysis:.....	4
Objective of the project:.....	4
Data Manipulation and Data Exploration: .....	5
New Column Names .....	5
Unique Values.....	6
Most Frequent Values .....	6
Data Visualization:.....	7
User Name.....	8
User Location.....	9
Tweet Source.....	9
Text Wordclouds .....	10
Top Hashtags.....	12
Setimental Analysis: .....	13
Conclusions and Recommendations: .....	16
Work Cited:.....	17

## Introduction:

Rafa Nadal is one of the greatest professional tennis players in history, he has won 22 Grand Slam titles and numerous other tournaments throughout his present career. When he won his most recent Grand Slam, Roland Garros in Paris, his fan base was quick to show their support on social media, especially on Twitter. In this project, I will be making a sentimental analysis on the tweets about Rafa Nadal's 22nd Grand Slam victory in order to understand the sentiments and emotions expressed by his fans. This analysis will provide a valuable understanding of how fans perceive and react to Nadal's accomplishments. Additionally, in the project report, I will discuss how this understanding of Nadal's fan base can help in revenue generation and contribute to the growth of the business of a tennis club. With a large and devoted fan base like Nadal has, there are a lot of opportunities for sponsorship deals, merchandise sales, and other revenue sources. Also, a strong fan base can lead to increased ticket sales and interest in attending Nadal's matches, which can help as well the grow of the club and the sport of tennis. By understanding the sentiment of Nadal's fan base, the club can use the sentiment analysis to improve their marketing and engagement with Nadal's fans.

## Problem Analysis:

In this project, I want to analyze the sentiment of the tweets related to Rafael Nadal's 22nd Grand Slam win. One of the problems this project has is to address the understanding of the emotional response of Nadal's fans to his achievement. This understanding can help the tennis club in adjusting their marketing efforts and engagement strategies to better connect with Nadal's fans and strengthen their relationship.

## Objective of the project:

The main objective with the sentimental analysis is to know how this will help to improve the tennis club business in terms of marketing, sponsorships, and customer loyalty. Another purpose of the project is to analyze the sentiments of Rafa Nadal's fans towards his 22nd Grand Slam win. By examining the tweets related to the event, we can understand how his fans feel about this achievement.

## Data Manipulation and Data Exploration:

```
In [2]: df = pd.read_csv("https://raw.githubusercontent.com/Smurkio8/FinalProject720/main/rafaelnadal_tweets.csv")
In [3]: df.head()
```

Out[3]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	h
0	Nong Nhat Minh	NaN	NaN	2021-11-26 07:07:47	20	890	1035	False	2022-06-08 17:02:44	@DappCensus Nice project. @linhair8 @LongAindr...	['dapp', 'gives
1	Peter Ndoro	Africa	Broadcast Journalist   This is not a News Feed...	2009-03-22 16:29:58	279853	191548	5787	True	2022-06-08 16:52:04	The champions are being born everyday. They ar...	
2	Gurpreet Singh	Mansa	https://t.co/2zAmCdu2Jh	2019-05-17 16:33:12	61	1214	2727	False	2022-06-08 16:43:24	@DappCensus 🚀 In Successful in 2022InBig profit...	
3	👉 Earning Tips 👉	Dhaka, Bangladesh	ARKERARMY 🙌	2020-08-28 08:56:58	115	2195	3716	False	2022-06-08 16:39:26	@DappCensus This is very huge and great projec...	
4	ahs	universe	a common man.	2012-06-08 09:23:24	35	393	21	False	2022-06-08 16:35:21	@neeteshb @RajKumarMUFC @87vintage @nadalprop ...	['T

First, we are going to install and import all the necessary libraries needed for the project and read the dataset to be able to analyze it. Thanks to the info() command we see that there are some values missing.

```
Missing Values

In [5]: data= df[df['user_location'].notna()]
data.shape

Out[5]: (6204, 13)
```

```
In [6]: def missing_data(df):
total = df.isnull().sum()
percent = (df.isnull().sum()/df.isnull().count()*100)
tt = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
types = []
for col in df.columns:
    dtype = str(df[col].dtype)
    types.append(dtype)
tt['Types'] = types
return(np.transpose(tt))

In [7]: missing_data(df)
```

Out[7]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source
Total	0	2555	936	0	0	0	0	0	0	0	1772	0
Percent	0.0	29.169997	10.686151	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.23062	0.0
Types	object	object	object	object	int64	int64	int64	bool	object	object	object	object

## New Column Names

In this process I will be renaming the column names to make it easier and to take some unnecessary information out in some of the column names.

## Rename of Columns

```
In [14]: df.columns = ['Name', 'Location', 'Description', 'Created_in', 'Followers', 'Friends', 'Favourites', 'User_verified', 'Date', 'Text', 'Hashtags', 'Source', 'Is_retweet']
df.head(2)
```

```
Out[14]:
```

	Name	Location	Description	Created_in	Followers	Friends	Favourites	User_verified	Date	Text	Hashtags	Source	Is_retweet
0	Nong Nhat Minh	NaN	NaN	2021-11-26 07:07:47	20	890	1035	False	2022-06-08 17:02:44	@DappCensus Nice project. @linhair8 @LongAirdr...	['dappcensus', 'Airdrop', 'BNB', 'giveaway', '...	Twitter Web App	False
1	Peter Ndoro	Africa	Broadcast Journalist   This is not a News Feed...	2009-03-22 16:29:58	279853	191548	5787	True	2022-06-08 16:52:04	The champions are being born everyday. They ar...	NaN	Twitter for iPhone	False

## Unique Values

Now I examine the unique values, because it helps to identify any outliers or anomalies that may impact the analysis. It also helps to identify any missing or incorrect values in the dataset.

### Unique Values

```
In [9]: def unique_values(df):
total = df.count()
tt = pd.DataFrame(total)
tt.columns = ['Total']
uniques = []
for col in df.columns:
    unique = df[col].nunique()
    uniques.append(unique)
tt['Uniques'] = uniques
return(np.transpose(tt))
```

```
In [10]: unique_values(df)
```

```
Out[10]:
```

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	i
Total	8759	6204	7823	8759	8759	8759	8759	8759	8759	8759	6987	8759	
Uniques	5568	2217	5033	5633	1954	1959	4635	2	7693	8701	1906	19	

## Most Frequent Values


Showing the most frequent values in the data set helps to have an idea of the distribution of the data and the range of values. It also helps in identifying the outliers, as they may have low frequencies or not appear in the most frequent list.

## Most Frequent Values

```
In [15]: def most_frequent_values(df):
total = df.count()
tt = pd.DataFrame(total)
tt.columns = ['Total']
items = []
vals = []
for col in df.columns:
    itm = df[col].value_counts().index[0]
    val = df[col].value_counts().values[0]
    items.append(itm)
    vals.append(val)
tt['Most frequent item'] = items
tt['Frequency'] = vals
tt['Percent from total'] = np.round(vals / total * 100, 3)
return(np.transpose(tt))
```

In [16]: most\_frequent\_values(df)

Out[16]:

	Name	Location	Description	Created_in	Followers	Friends	Favourites	User_verified	Date	Text	Hashtags	Source	Is_retweet
Total	8759	6204	7823	8759	8759	8759	8759	8759	8759	8759	6987	8759	8759
Most frequent item	 PS Junsu / hana	India	JYJBB BTS BARCA #AOT @fcbarcelona @RafaelNadal...	2012-02-07 04:25:29	6	54	2	False	2022-06-05 15:35:40	I swear to God, if you knew the secret we are ...	[RafaelNadal%]	Twitter for Android	False
Frequency	80	314	80	80	125	83	88	8517	9	8	2121	3946	8759
Percent from total	0.913	5.061	1.023	0.913	1.427	0.948	1.005	97.237	0.103	0.091	30.356	45.051	100.0

In [17]: print(f"data shape: {df.shape}")

data shape: (8759, 13)

## Data Visualization:

Now in the data visualization I will be showing the count of the most common names, locations, and tweet sources using the same code.

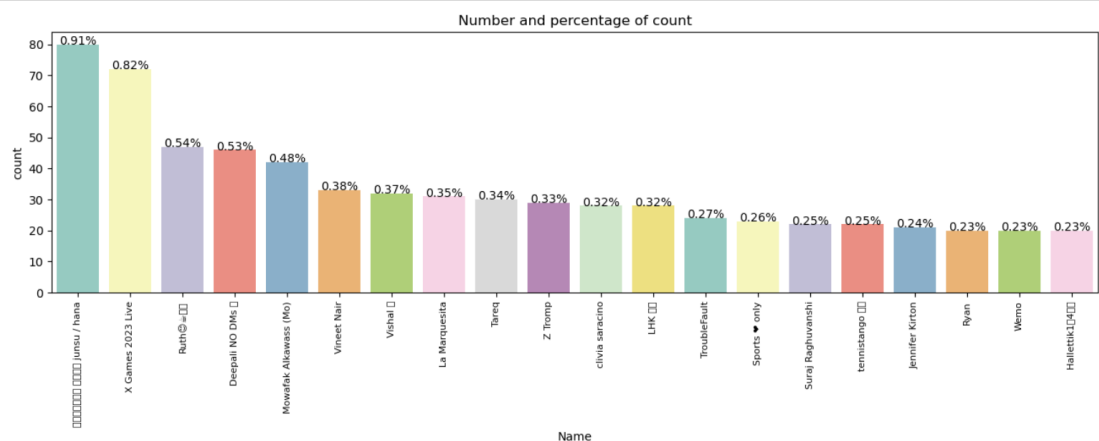
## Data Visualization

```
In [19]: def plot_count(feature, title, df, size=1, ordered=True):
    f, ax = plt.subplots(1,1, figsize=(4*size,4))
    total = float(len(df))
    if ordered:
        g = sns.countplot(x=feature, data=df, order = df[feature].value_counts().index[:20], palette='Set3')
    else:
        g = sns.countplot(x=feature, data=df, palette='Set3')
    g.set_title("Number and percentage of {}".format(title))
    if(size > 2):
        plt.xticks(rotation=90, size=8)
    for p in ax.patches:
        height = p.get_height()
        ax.text(p.get_x()+p.get_width()/2.,
                height,
                '{:1.2f}%'.format(100*height/total),
                ha="center")
    plt.show()
```

## User Name

### User Name

```
In [20]: plot_count("Name", "count", df,4)
```



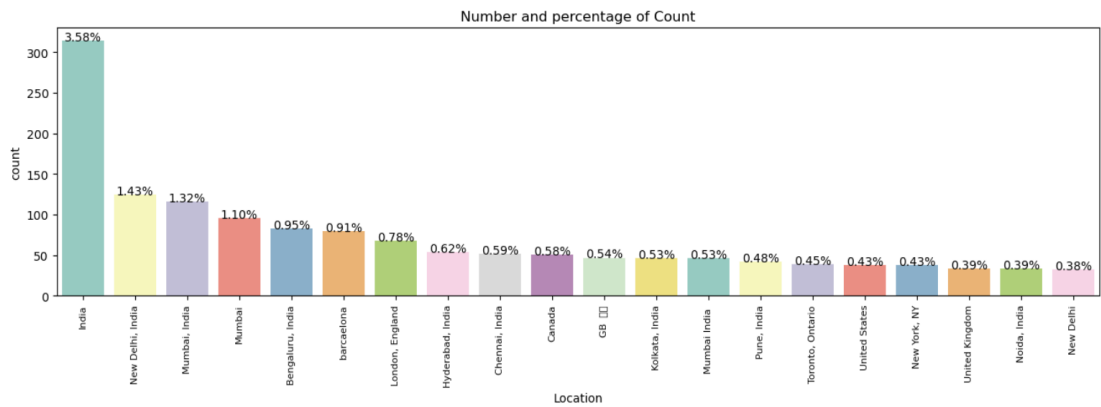
We can observe that a lot of names are originally Indian and then there are some other common names that have being created for the love of the sport as “Sports only” or “tennistango”.



## User Location

### User Location

```
In [21]: plot_count("Location", "Count", df,4)
```

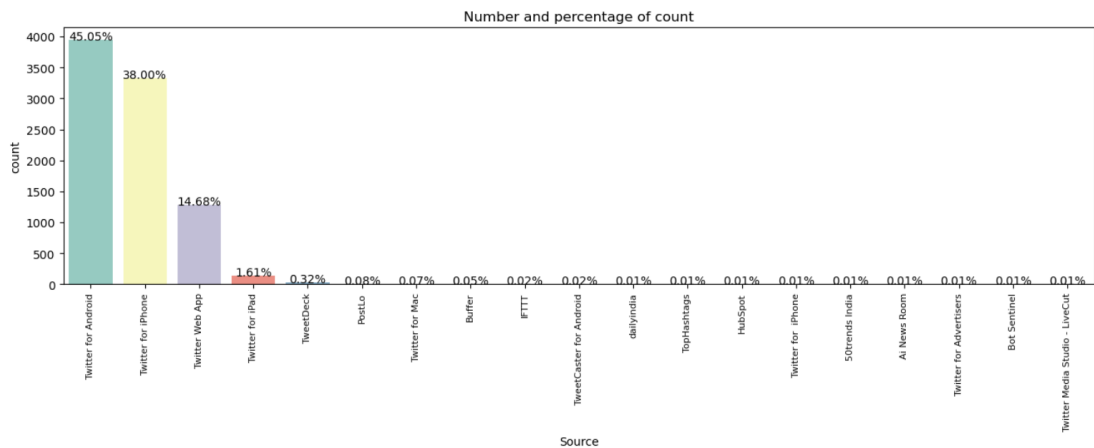


As I mentioned in the plot before, a lot of common names are coming from India and that is because the user location we can observe how India has a much bigger percentage of the count in the graph. In this chart India is clearly the winner followed by Barcelona, United Kingdom, Canada, and United States.

## Tweet Source

### Tweet Source

```
In [22]: plot_count("Source", "count", df,4)
```



We can observe that the most tweet source used by the users are clearly Twitter for Android and Twitter for iPhone. This may also be a reference on what is the phone market dominance in the countries listed above.

## Text Wordclouds

A wordcloud is a visual representation of text data where the size of each word represents its frequency in the dataset. In this project, it will represent the tweets related to Rafael Nadal's 22nd Grand Slam Title, so the wordcloud will give us a quick overview of the most commonly used words in those tweets.

By using a tennis ball as the shape for the wordcloud, I add an additional layer of meaning to the visualization, which it will be the sport "Tennis".

### Text Wordclouds

```
In [23]: mask = 255 - np.array(Image.open('tennis-ball-.png'))
```

```
In [24]: def show_wordcloud(data, title="", mask=None, color="white"):
    text = " ".join(t for t in data.dropna())
    stopwords = set(STOPWORDS)
    stopwords.update(["t", "co", "https", "amp", "u", "Rafa", "rafaelnadal", "Rafaelnadal", "Nadal", "FrenchOpen", "RolandGarros"])
    wordcloud = WordCloud(stopwords=stopwords, scale=4, max_font_size=50, max_words=500, mask=mask, background_color=color).generate(text)
    fig = plt.figure(1, figsize=(16,16))
    plt.axis('off')
    fig.suptitle(title, fontsize=20)
    fig.subplots_adjust(top=1.0)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.show()
```

```
In [25]: show_wordcloud(df['Text'], title = 'Prevalent words in tweets', mask=mask)
```

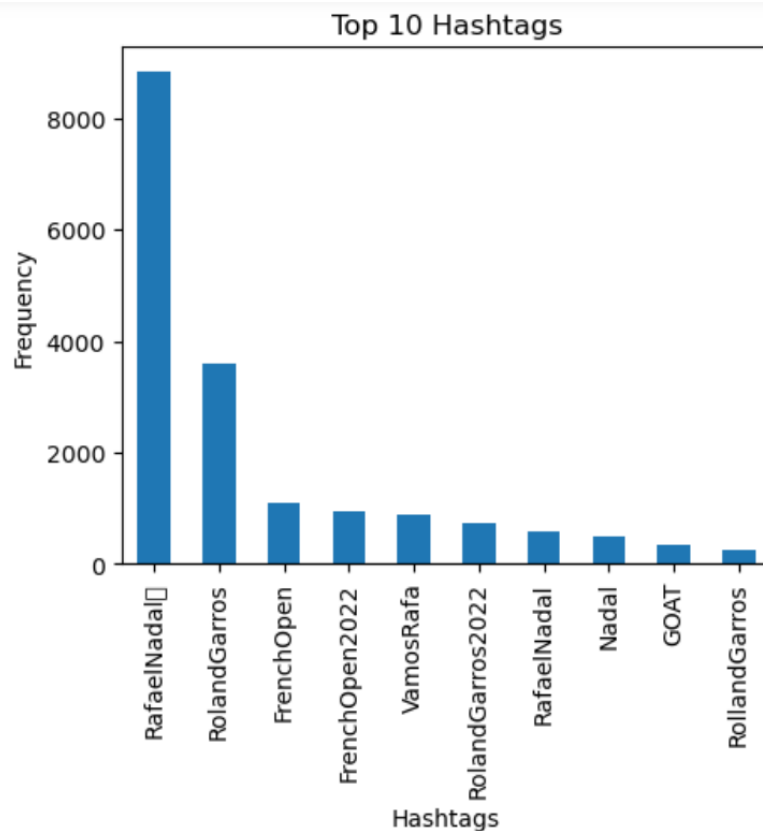
[illegible]



## Top Hashtags

```
hashtags = []
for tweet in df.Text:
    hashtag = re.findall(r"#(\w+)", tweet)
    hashtags.extend(hashtag)

# Count the frequency of each hashtag and plot the top 10
plt.figure(figsize=(5, 4))
top_hashtags = pd.Series(hashtags).value_counts().head(10)
top_hashtags.plot(kind='bar')
plt.title("Top 10 Hashtags")
plt.xlabel("Hashtags")
plt.ylabel("Frequency")
plt.show()
```



## Sentimental Analysis:

Sentiment analysis is a powerful tool for understanding how people feel about a particular topic, brand, or person. In this case, I am referring to Rafael Nadal's

22nd Grand Slam Title, and this sentimental analysis can help us understand how people felt about the tennis legend during his achievement and how this can improve the business of the tennis club in terms of marketing.

By doing the sentimental analysis of the tweets related to this topic, we can have a better understanding about how people perceive Rafa Nadal as a player, his performance on the final, and his overall reputation in the tennis community.

## Sentimental Analysis

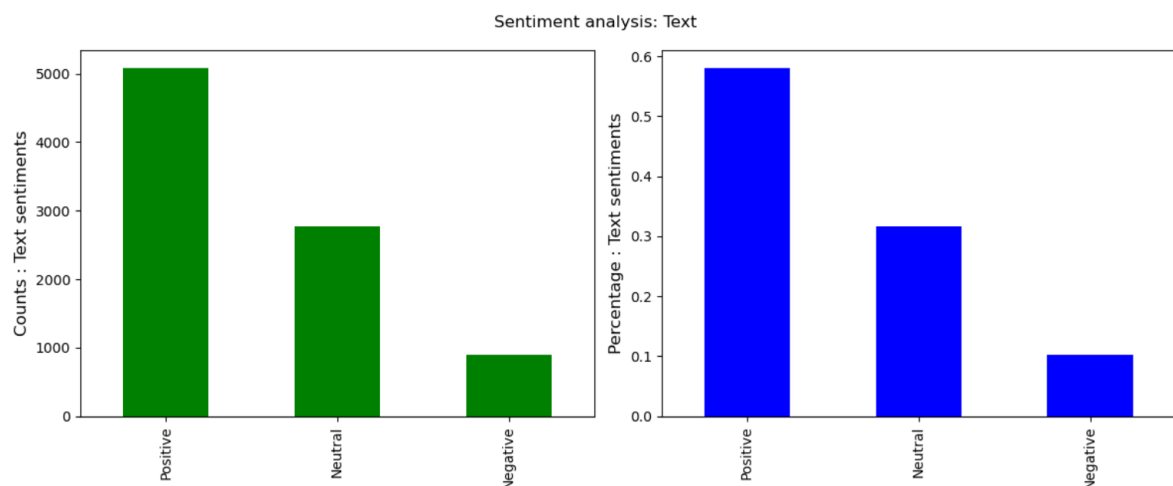
```
warnings.simplefilter("ignore")
sia = SentimentIntensityAnalyzer()
def find_sentiment(post):
    try:
        if sia.polarity_scores(post)["compound"] > 0:
            return "Positive"
        elif sia.polarity_scores(post)["compound"] < 0:
            return "Negative"
        else:
            return "Neutral"
    except:
        return "Neutral"
```

```
def plot_sentiment(df, feature, title):
    counts = df[feature].value_counts()
    percent = counts/sum(counts)

    fig, (ax1, ax2) = plt.subplots(hcols=2, figsize=(12, 5))

    counts.plot(kind='bar', ax=ax1, color='green')
    percent.plot(kind='bar', ax=ax2, color='blue')
    ax1.set_ylabel(f'Counts : {title} sentiments', size=12)
    ax2.set_ylabel(f'Percentage : {title} sentiments', size=12)
    plt.suptitle(f"Sentiment analysis: {title}")
    plt.tight_layout()
    plt.show()
```

```
df['text_sentiment'] = df['Text'].apply(lambda x: find_sentiment(x))
plot_sentiment(df, 'text_sentiment', 'Text')
```



The plot shows the sentiment analysis results for the text data in the 'Text' column of the dataset. The plot is divided into two subplots: the left subplot shows the counts of the sentiment labels (positive, negative, and neutral), and the right subplot shows the percentages of the sentiment labels.

Based on the plot, it is clear that the sentiment of the tweets related to Rafael Nadal's achievement was more positive than neutral, and it was barely negative. The majority of the tweets were classified as positive, with a smaller number of tweets classified as neutral or negative.

The exact counts and percentages of the sentiment labels can be read from the y-axis of the plot. For example, if we look at the left subplot, we can see that there were around 5000 tweets classified as positive, around 2900 tweets classified as neutral, and around 900 tweets classified as negative.

Overall, the sentiment analysis results suggest that the public perception of Rafael Nadal during his 22nd Grand Slam final was largely positive, with relatively few negative or neutral tweets. These results could be used to gain insights into how the public perceives Nadal as a player, his performance during the final, and his overall reputation in the tennis community.

### Conclusions and Recommendations:

The conclusion based on the whole analysis and the sentimental analysis is that the public sentiment towards Rafael Nadal during his 22nd Grand Slam Title was overall positive. This could potentially be an opportunity of growth in the tennis club business to attract more customers or improve customer loyalty. These are a few ways how the sentimental analysis could potentially improve the business:

**Marketing:** The positive sentiment towards Nadal could be used in marketing campaigns to attract more customers to the tennis club. For example, the club could advertise itself as a place where fans of Nadal can come to watch matches and discuss his performance or even watch him playing.

**Sponsorship opportunities:** If the tennis club is interested in sponsoring players or tournaments, the sentiment analysis results could be used to identify players who are popular among fans. For example, if there is a player who has a similarly positive sentiment score to Nadal, sponsoring that player or tournament could be a good way to align the club's brand with positive public sentiment.



Overall, the sentiment analysis results provide valuable insights into how the public perceives Rafael Nadal and could potentially be used to inform marketing strategies or sponsorship decisions for your tennis club business.

Work Cited:

<https://www.kaggle.com/code/ankanhore545/rafael-nadal-tweets/notebook>