

COMP 541
California State University, Northridge
Fall 2022
Final Project Report

Analyzing Twitter Population Sentiments
based on LA Lakers Wins and Losses

Assignment Submitted By:- Group 1

- FNU Hardik
- Kumar Sai Reddy Guntaka
 - Sravya Guntupalli
- Abinay Kumar Gundu
 - Raj Kumar Lakoji

Table Of Contents

S.no	Title	Page No.
1	Business Understanding	3
2	Determining Business Objectives	3
3	Assessing Situation	4
4	Determine Data Mining Goals	6
5	Data Understanding	7
6	Dataset: Details	7
7	Dataset: Description	9
8	Data Preparation	11
9	Data Exploration	11
10	Data Visualization	14
11	Data Cleaning and Transformation	17
12	Feature Selection	19
13	Feature Extraction	22
14	Modeling	25
15	Evaluation	26
16	Direct Approach	27
17	Indirect Approach	29
18	Deployment	32
19	Summary of Learning Experiences	33
20	References	34
21	Appendix	35

Business Understanding

Determining Business Objectives

Background :- Big Companies that decide to sponsor sports teams or when they want to invest in a new venture out of their realm of expertise will often try different techniques to analyze the feasibility of the new ventures. Our project is an attempt to provide an efficient and modular system that can help companies with a variety of market research related tasks.

This analysis allows us to quantify the sentiment of the LA Lakers fans based on the NBA games the LA Lakers win or loss. This system can be extended to quantify and analyze the behavior of the population on a new public policy or a major event [1].

Business Objectives :- The business objective is as follows: To help investors determine whether to sponsor any particular team based on how the general public feels about them or to estimate sales by tracking sentiment of the twitter population towards the product.

This can also be done to deliberate on new product launches or to determine how current geo-political scenarios are perceived by the general twitter population.

Investors or sponsors can use our sentiment analyzed data to determine whether to proceed with their investment or to look at different avenues for investments.

Business Success Criteria :- The success criteria for us will be to be able to successfully predict how the overall feeling / sentiment of the general public is towards any sports team or any particular product, so investors can have reliable information to then make decisions.

By identifying the users' emotions, you can get a better idea of their experience and provide better customer service, which eventually leads to a decrease in customer churn. A brand wants to know when the fans are most likely to purchase the product when sponsored.

Monitor and measure customer satisfaction and perception during a certain marketing campaign by utilizing sentiment analysis and associating customer attitudes with product launches to further adjust either on the campaign or the product [2].

Assessing Situation

Inventory Of Resources :- The resources available to us for this project include 5 data miners (in-training), twitter data which will be gathered from twitter live during the project testing and implementation, Dell workstations, Jupyter notebook and anaconda package manager which are the platforms on which we'll implement and run our project.

Requirements, Assumptions and Constraints :- The required data that we will run our language processing algorithm on will always be present online and on a social media site like twitter all the data is public and hence can be used without any permission.

Assumptions regarding our project include that people actually feel how they express themselves on social media. If a large amount of twitter population says one thing on twitter and then goes on to do the opposite thing in real-life then our analysis might generate erroneous results.

Some constraints we might encounter could be in the form of the size of the dataset we gather from the twitter hashtags we search for. In order to keep our project resource utilization low we will only scrape 2000 or so tweets from the internet when we run our model.

We will scrape Twitter tweets using a python library called tweepy. Before you start using tweepy, you would need a Twitter developer account to call Twitter APIs. We will get access after some time. We need 4 pieces of information ready- API Key, API secret key, Access token, and access token secret [3].

Risks and Contingencies :- We are currently experimenting on different ways in which we can gather data from twitter handles in real time. We are currently exploring the tweepy library of python functions to accomplish this task.

Contingency plans for if we are not able to gather live data at run time for our model would include using already available datasets to perform sentiment analysis onto.

Terminology :- Relevant business terms :

Investor – a person or organization that puts money into financial plans, property, etc. with the expectation of achieving a profit.

Sponsor – an individual or organization that pays some or all of the costs involved in staging a sporting or artistic event in return for advertising.

Merchandise - promote the sale of (goods), especially by their presentation in retail outlets.

Sentiment - a view of or attitude toward a situation or event; an opinion.

Data Mining terminology :

Data Source - A data source is simply the source of the data. It can be a file, a particular database on a DBMS, or even a live data feed. The data might be located on the same computer as the program, or on another computer somewhere on a network.

Data Cleaning - Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Sentiment Analysis - the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Correlation Analysis - Correlation Analysis is statistical method that is used to discover if there is a relationship between two variables/datasets, and how strong that relationship may be.

Costs and benefits :- Potential benefits to any company or individual looking to perform sentiment analysis on the twitter population could include an increase in sales if the target of analysis is a product, or potential benefits could include guiding bets on sports teams or figures based on public sentiment. The costs associated with creating this project are estimated to be about 200 engineering hours and then you only need 5 minutes to change the focus of the sentiment analysis by changing the twitter handles and hashtags from which to scrape the data.

Determine Data Mining Goals

Data Mining Goals :- Intended goal of this project will be an overall estimation of whether the twitter population feels positively, negatively, or neutrally about any particular twitter hashtag.

Data Mining Success Criteria :- If we can manage to get a high correlation of positive sentiments with recent positive events related to the entity being analyzed we will know that our model works well in real time scenarios and can be a sign of success for our project. (For example:- If we are analyzing twitter population sentiments related to a recent game and assuming that the particular team wins the game, then a highly positive sentimental analysis result would indicate that our model works well) [4].

This is an unstructured dataset scenario, neither are we doing classification or regression so classical performance measures are not useful here, but if I am to assign labels beforehand then I can use performance measures like precision, recall that are conventionally used to assess model performance.

Initial Assessment of Tools and Techniques :- As per our initial assessment we have identified the following python libraries and API's that will help us in completing our project, they are:

Tweepy : Tweepy is an open-sourced, easy-to-use Python library for accessing the Twitter API. It gives you an interface to access the API from your Python application.

Numpy : NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project, and you can use it freely. NumPy stands for Numerical Python.

Matplotlib : Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc.

Pandas : Pandas is an open-source library in Python. It provides ready to use high-performance data structures and data analysis tools. Pandas module runs on top of NumPy, and it is popularly used for data science and data analytics.

Data Understanding

Dataset: Details

Acquired Dataset :- Shown below is a screenshot of the dataset we collected. The dataset was collected real time from twitter and saved as a .csv file after being stored as a data-frame inside our Jupyter Notebook.

	A	B	C	D	E	F
1	User	Tweet	Date			
2	0 Lakers	Before there was Season 20, there was Season 1.	2022-09-26 02:10:00+00:00			
3	1 Lakers	Reconciliation, redemption, and recognition.	2022-09-25 23:59:00+00:00			
4	2 Lakers	The Los Angeles Lakers have requested waivers on forward Fabian White Jr.	2022-09-25 20:25:12+00:00			
5	3 Lakers	â€œI donâ€™t know who was under the rim, but you shouldâ€™ve known not to jump” ðŸŒŸ	2022-09-25 18:36:51+00:00			
6	4 Lakers	How would Thomas Bryant describe his game?	2022-09-24 07:00:01+00:00			
7	5 Lakers	It’s about the work, always about the work âšŒij, https://t.co/H6ogMjOGi5	2022-09-23 18:13:00+00:00			
8	6 Lakers	Movies, music, & more with a new face on the squad.	2022-09-22 17:31:00+00:00			
9	7 Lakers	A full circle moment.	2022-09-22 02:02:44+00:00			
10	8 Lakers	EVERY LeBron dunk as a Laker	2022-09-21 23:34:00+00:00			
11	9 Lakers	5x Champion, All-Star, MVP, and the man behind the Mikan Drill.	2022-09-21 18:59:00+00:00			
12	10 Lakers	Step by step	2022-09-20 20:31:00+00:00			
13	11 Lakers	ðŸŒœðŸŒš https://t.co/aWpEGvsURv	2022-09-20 17:16:56+00:00			
14	12 Lakers	Hook shot forever https://t.co/olxPAlsp5	2022-09-20 17:07:15+00:00			
15	13 Lakers	Like they never stopped: Showtime â€˜80s Å» 2022 https://t.co/tF9CHpyfDp	2022-09-20 16:56:58+00:00			
16	14 Lakers	RT @Dodgers: Welcome to LA, Coach Ham! https://t.co/bOmFQV8h7	2022-09-20 03:18:11+00:00			
17	15 Lakers	Coach Darwin took to the field to throw out tonightâ€™s first pitch in celebration of Black Heritage Night	2022-09-20 03:04:05+00:00			
18	16 Lakers	Play ball!	2022-09-20 02:43:51+00:00			
19	17 Lakers	A dynasty reunited.	2022-09-19 22:04:43+00:00			
20	18 Lakers	Your voice counts.	2022-09-19 21:18:51+00:00			
21	19 Lakers	We’re talking mentors turned teammates, favorite kicks, & more.	2022-09-19 20:47:00+00:00			
22	20 Lakers	Dunks, dimes, diving on the floor – Dennis is bringing back a lot of heart & hustle. https://t.co/fdiw	2022-09-19 18:01:00+00:00			
23	21 Lakers	RT @champagnennuts: STILL an unstoppable shotâ€™ at 75â€¦ https://t.co/rQ0c7dsp6a	2022-09-19 05:56:01+00:00			
24	22 Lakers	The grind got to them.	2022-09-19 00:15:00+00:00			
25	23 Lakers	Be great https://t.co/j4DqZL8TV	2022-09-18 23:13:00+00:00			
26	24 Lakers	Posters, pre-game meals, and seeing that Lakers jersey for the first time.	2022-09-18 18:15:00+00:00			
27	25 Lakers	RT @JamesWorthy42: #ShowtimeReunion Roster 2022 https://t.co/bf7uEUfmlc	2022-09-18 17:37:18+00:00			
28	26 Lakers	Weekend work https://t.co/bE6NTdyQeQ	2022-09-17 20:12:00+00:00			
29	27 Lakers	Forever a Champion, Happy Birthday Phil Jackson â–, https://t.co/OdMboYNH4r	2022-09-17 16:07:00+00:00			
30	28 Lakers	OFFICIAL: That’s Tuff ðŸŒšðŸŒš https://t.co/KgxrU527U1	2022-09-17 00:40:31+00:00			
31	29 Lakers	Showtime Reunion in Maui ðŸŒœ https://t.co/FXBuClxos	2022-09-16 21:29:07+00:00			

Figure 1: Dataset Collected from twitter using Tweepy API

Location Of Dataset :- This Dataset was acquired by utilizing the Twitter API library and the functions provided within it. The Focus for gathering this data was mostly twitter handles that are closely associated with the LA Lakers Basketball team.

Method of Acquisition :- This dataset was acquired by utilization of the tweepy library of specific twitter related python functions [5-8]. Such access to twitter data is controlled by Twitter and you must first apply for an elevated developer account in order to be able to work with the twitter functions that will be used to read the data and the account that will be used to access data online will be via the same developer account. Meaning any data made public by unfollowed twitter handles will be readily available for our purposes. Creating and being able to use the twitter elevated developer account will require at least 48 hours (2 days) as twitter has certain security measures and verification procedures it must do before we are granted access.

For our data mining purpose we have gathered '1000' tweets from 11 different twitter handles that are dedicated to news and current affairs related to the LA Lakers team. Within our code the data gathering works in the following manner:

1. We create an empty list to store any data we might append into it.
2. We use another empty list and fill it up with '1000' of the latest tweets made by the targeted twitter handle.
3. Then we take each item from the list used in step 2, separate it into 3 parts: User, Tweet and Date and store it into a variable x.
4. One at a time we append the variable x to the empty list from step 1 to create the full dataset.
5. But the dataset at this point is unorganized and is not even separated into different lines (see output of block 5).
6. Using the Pandas library the data from step 5 is neatly organized into a dataframe.
7. Using the Pandas library again we can convert this dataframe into a .csv file easily

Problems Encountered :- During this part of the data mining project we encountered many problems, some of the major problems are listed below:

1. Getting the correct Access tokens and Consumer access tokens needed for reading real time twitter data using tweepy API.

(Solution) The solution for this problem was found easily by watching YouTube videos on how to read real time data from twitter using the tweepy API.

2. Even after we got the correct access tokens, we didn't realize that we needed to upgrade our developer from essential developer access to elevated developer access.

(Solution) This process was straight-forward as all we had to do was fill out a few forms for twitter explaining what we wish to do with the data we will gather and then wait for 48 hours for clearance.

3. The data that is parsed from Twitter will be unorganized and will not resemble any dataframe we have previously worked with.

(Solution) This was a simple solution, as explained earlier in the acquisition method part, we can use empty lists to store data elements one by one and then use the pandas library to convert it into an organized dataframe.

Dataset: Description

Format of Data :- The data format for our dataset is text files. Any data parsed from twitter like tweets, pictures, date (numeric data), usernames are all converted into text format when given out to the empty list waiting to receive this data, this can cause special characters and emojis and other special icons to look unrecognizable or appear as gibberish text in our dataset, this has to be removed when cleaning the dataset later.

Quantity of Data :- For our project we have collected '1000' tweets from '11' different twitter handles that are most closely associated with the LA Lakers basketball team. Our dataframe consists of 3 columns: Username, Tweet (actual tweet in text form), Date. The size of dataframe after being converted into .csv format is around 1,510 Kb.

We were limited to this size of almost 11,000 tweets for a number of reasons:

- The twitter API needs some downtime between each successive read when reading data from twitter handles, this is recommended to be kept at 3 seconds otherwise we were seeing errors in retrieving data as the Twitter API places restrictions on how much data you are allowed to retrieve within

a set period of time. Hence the time required to gather more than 11,000 tweets increases drastically.

- Later on we have to perform data cleaning as well onto all these data instances, this will become more complex and resource heavy for the computer as we increase the number of instances, hence, to try and devise an optimal analyzer model we have limited the data gathering to 11,000 samples.

Identify the Type of Data :- The data we have gathered is of a semi-structured nature, and when we gather the data, we are not aware of the sentiment that is attached to each tweet. We collected the unstructured data from Twitter using the tweepy library. We are using methods like OAuthHandler, under the tweepy library and storing the data unstructured data in a list called gathered tweets.

Identify the Type of Dataset :- Our dataset is of text (web page) data type as we have gathered it all from twitter in real time and it is provided to us in textual format by the twitter API.

Identify the Attribute Types :- Our dataset contains three columns or three major attributes, their types are discussed below:

User – Categorical, because the user handles from which to gather data are provided by us in the form of a “users” list, hence each username belongs to one of the categories from the list we gave the twitter API. It is basically a nominal attribute consists of data from user list.

Tweet – Textual, because we know from the twitter API documentation that any actual tweet content be it numbers, emoticons or actual text is all combined into a textual format when parsed from the web.

Date – Ordinal because date has an inherent order associated with its numbering sequence. For example:- the newer date being higher in order from the older date.

Data Preparation

Data Exploration

First Look at Gathered Data :- Since we gathered our data from twitter i.e., a social media platform, the dataset initially is in a semi-structured format. As can be seen in the image below.

```
print ((gathered_tweets))

[['Lakers', 'Before there was Season 20, there was Season 1. \n\n@KingJames reflects on his first points in the league 🏀 ht
tps://t.co/o01cxuBCMF', datetime.datetime(2022, 9, 26, 2, 10, tzinfo=datetime.timezone.utc)], ['Lakers', 'Reconciliation, red
emption, and recognition.\n\nStream the next chapter of the #LakersDoc Monday on @Hulu. https://t.co/2sbM7Kkbs', datetime.da
tetime(2022, 9, 25, 23, 59, tzinfo=datetime.timezone.utc)], ['Lakers', 'The Los Angeles Lakers have requested waivers on forw
ard Fabian White Jr.', datetime.datetime(2022, 9, 25, 20, 25, 12, tzinfo=datetime.timezone.utc)], ['Lakers', '"I don't know w
ho was under the rim, but you should've known not to jump" 🌟\n\n@LakeShow | @lonniewalker_4 https://t.co/Uy6G27NoVq', dateti
me.datetime(2022, 9, 25, 18, 36, 51, tzinfo=datetime.timezone.utc)], ['Lakers', 'How would Thomas Bryant describe his game?
\n\n"Very intense." https://t.co/F4P00v9HdY', datetime.datetime(2022, 9, 24, 7, 0, 1, tzinfo=datetime.timezone.utc)], ['Laker
s', '"It's about the work, always about the work 🌟 https://t.co/H6ogMj0Gi5", datetime.datetime(2022, 9, 23, 18, 13, tzinfo=da
atetime.timezone.utc)], ['Lakers', 'Movies, music, & more with a new face on the squad.\n\nGet to know @troy_brown33 http
s://t.co/vWicK0kwbz', datetime.datetime(2022, 9, 22, 17, 31, tzinfo=datetime.timezone.utc)], ['Lakers', 'A full circle momen
t. \n\nCoach Darwin got it goin' for the Dodgers Monday night. https://t.co/VQJ4H06hL", datetime.datetime(2022, 9, 22, 2, 2,
44, tzinfo=datetime.timezone.utc)], ['Lakers', 'EVERY LeBron dunk as a Laker\n\n...so far 🌟', datetime.datetime(2022, 9, 21,
23, 34, tzinfo=datetime.timezone.utc)], ['Lakers', '5x Champion, All-Star, MVP, and the man behind the Mikan Drill.\n\nOn Oct
ober 30th we celebrate Mr. Basketball. https://t.co/ydUOggboFQ', datetime.datetime(2022, 9, 21, 18, 59, tzinfo=datetime.timez
one.utc)], ['Lakers', 'Step by step\nRep by rep https://t.co/UdYmNhwFRj', datetime.datetime(2022, 9, 20, 20, 31, tzinfo=datet
ime.timezone.utc)], ['Lakers', '💖 https://t.co/awpEGvsURV', datetime.datetime(2022, 9, 20, 17, 16, 56, tzinfo=datetime.ti
mezone.utc)], ['Lakers', 'Hook shot forever https://t.co/olxtPAJsp5', datetime.datetime(2022, 9, 20, 17, 7, 15, tzinfo=dateti
me.timezone.utc)], ['Lakers', 'Like they never stopped: Showtime '80s » 2022 https://t.co/tF9CHpyfDp', datetime.datetime(202
```

Figure 2: Initial Acquired Dataset

The next step will be to structure this gathered data set. But we still need to keep in mind that this data is not clean, this will contain all kinds of punctuations, digits, symbols, hyperlink addresses and emoticons because this is all unfiltered data.

Structuring the Gathered Data :- In the Image below you can see the 'Dataframe' Python function being used to read from the gathered data list and create a well-defined Dataframe.

```
In [6]: # Making an organized dataframe out of the list of tweets

data = []
for tw in gathered_tweets:
    for i in enumerate(tw):
        data.append(i[1])

tweet_data=pd.DataFrame(data=data, columns=['User', 'Tweet', 'Date'])
```

Figure 3: Converting Gathered Data (list) into a Dataframe

After this python cell, our dataset has been stored into a dataframe called 'tweet_data'. The image below shows the dataset that is now neatly transformed in a structured manner.

tweet_data

Out[9]:

	User	Tweet	Date
0	Lakers	Before there was Season 20, there was Season 1...	2022-09-26 02:10:00+00:00
1	Lakers	Reconciliation, redemption, and recognition.\n...	2022-09-25 23:59:00+00:00
2	Lakers	The Los Angeles Lakers have requested waivers ...	2022-09-25 20:25:12+00:00
3	Lakers	"I don't know who was under the rim, but you s...	2022-09-25 18:36:51+00:00
4	Lakers	How would Thomas Bryant describe his game? \n\...	2022-09-24 07:00:01+00:00
...
10685	LakersCommunity	#TreesforThrees https://t.co/9mYUOtD5l4	2015-10-29 03:17:38+00:00
10686	LakersCommunity	🎋 Thanks @NickSwagyPYoung! #TreesforThrees ht...	2015-10-29 03:08:40+00:00
10687	LakersCommunity	Opening up the season w/ great auction items: ...	2015-10-29 01:10:07+00:00
10688	LakersCommunity	#TreesforThrees is back! Each 3 pointer the @L...	2015-10-28 17:22:35+00:00
10689	LakersCommunity	Hey @lakers fans, only 5 more days till #Laker...	2015-10-23 16:10:06+00:00

10690 rows × 3 columns

Figure 4: Structured form of Gathered Dataset

Tidying up Gathered Data :- Our aim now is to make the dataset even clearer for any type of visualization or pre-processing that may be applied onto it later on. The main thing we need to take care of right now are special characters like emoticons, random https links in our tweet text, map symbols, iOS & Android flags etc. To accomplish this task we use the Python code shown in the image below.

```

In [11]: # Removing emoji, http link from text data - Data Cleaning

def remove(x):
    emoji_pattern = re.compile("[
        u"\U0001F600-\U0001F64F" # emoticons
        u"\U0001F300-\U0001F5FF" # symbols & pictographs
        u"\U0001F680-\U0001F6FF" # transport & map symbols
        u"\U0001F1E0-\U0001F1FF" # flags (iOS)
        u"\U00002500-\U00002BEF" # chinese char
        u"\U00002702-\U000027B0" # flags (Android)
        u"\U000024C2-\U0001F251" # positional/gps symbols
        u"\U0001F926-\U0001F937" # Emoji set 2
        u"\U00010000-\U0010ffff" # Emoji set 3
        u"\u2640-\u2642" # Emoji set 4
        u"\u2600-\u2B55" # Emoji set 5
        u"\u200d" # Zero width joiner
        u"\u23cf" # Eject symbol
        u"\u23e9" # black right-pointing double triangles
        u"\u231a" # Unicode escape
        u"\ufe0f" # dingbats
        u"\u3030" # wavy dash
    "]+", re.UNICODE)
    x = re.sub(emoji_pattern, '', x)
    x = re.sub(r'http\S+', '', x)
    return x

tweet_data['Tweet'] = tweet_data['Tweet'].apply(remove)
# print(tweet_data.head(10))

```

Figure 5: Removing miscellaneous characters from Gathered Data

The image below shows the dataset after all miscellaneous characters have been removed.

```

In [13]: # Total Size and Visual Check post Data Cleaning

tweet_data

```

	User	Tweet	Date
0	Lakers	Before there was Season 20, there was Season 1...	2022-09-26 02:10:00+00:00
1	Lakers	Reconciliation, redemption, and recognition.\n...	2022-09-25 23:59:00+00:00
2	Lakers	The Los Angeles Lakers have requested waivers ...	2022-09-25 20:25:12+00:00
3	Lakers	"I don't know who was under the rim, but you s...	2022-09-25 18:36:51+00:00
4	Lakers	How would Thomas Bryant describe his game? \n\...	2022-09-24 07:00:01+00:00
...
10685	LakersCommunity	#TreesforThrees	2015-10-29 03:17:38+00:00
10686	LakersCommunity	Thanks @NickSwagyPYoung! #TreesforThrees	2015-10-29 03:08:40+00:00
10687	LakersCommunity	Opening up the season w/ great auction items: ...	2015-10-29 01:10:07+00:00
10688	LakersCommunity	#TreesforThrees is back! Each 3 pointer the @L...	2015-10-28 17:22:35+00:00
10689	LakersCommunity	Hey @lakers fans, only 5 more days till #Laker...	2015-10-23 16:10:06+00:00

10690 rows x 3 columns

Figure 6: Clean form of Gathered Dataset

Data Visualization

Stop-Word Removal :- Stop words are common words like ‘the’, ‘and’, ‘is’, ‘for’ etc. that are very frequent in text, and so they do not present or offer any significant insight into the specific topic of the document. We must remove these stop words from the text in a given corpus to clean up the data, and easily identify words that are more rare and potentially more relevant to what we’re interested in [9].

The process for stop word removal starts by first tokenizing the words by using the nltk word_tokenize function. The list of stop words for removal is sourced from the NLTK English library. But we also add a few stop words from our side as they were appearing much more frequently than other words, they were “https” and “lakers”. Before ending the stop word removal stage we also manually remove all punctuations and digits from our dataset by using an iterative for loop and empty lists.

The frequency plots of the 20 most frequent words before and after the stop word removal are shown below in figure 7.

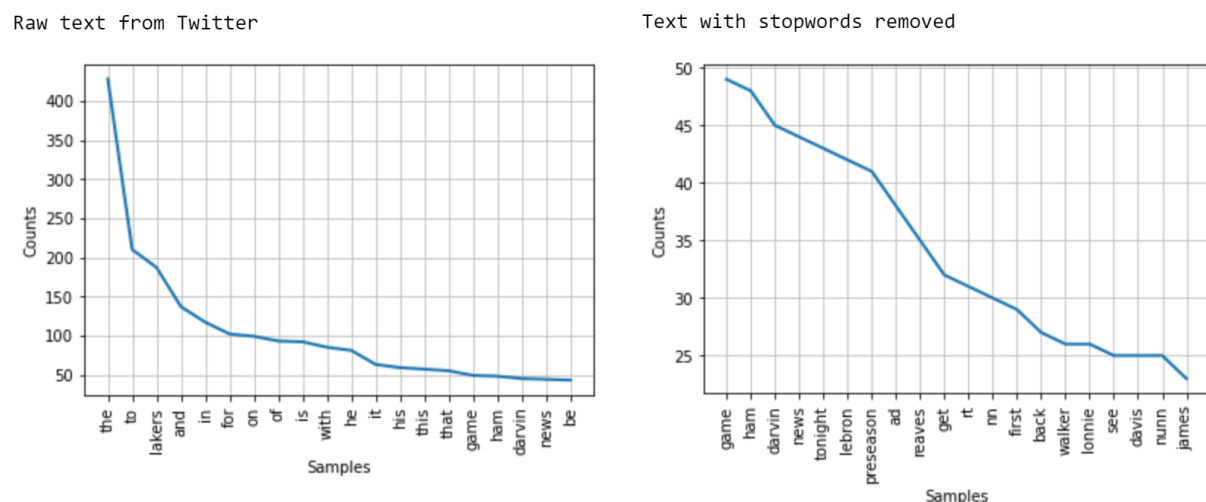


Figure 7: Frequency distribution before (left) and after (right) stop word removal

Before we continue further with the data visualization phase, we first need to talk about two more relevant visualizations that are available to us when we are trying to predict sentiment based off of textual data.

Now for any sentiment analysis model, we can have two basic approaches at classifying sentiments:

- The First is a Supervised machine learning or deep learning approach.
- The second is an Unsupervised 'Lexicon' based approach.

The data that we gather is sourced real time from Twitter and hence does not have any sentiment pre-associated with each tweet, in other words our dataset does not have a target feature, hence we cannot use supervised machine learning techniques.

For the Unsupervised Lexicon based approach we choose to use the 'VADER' (Valence Aware Dictionary and sEntiment Reasoner) lexicon library, and that in turn gives us access to two very useful visualization tools:-

1. Polarity Score
2. Subjectivity Score

Polarity Score: In essence, polarity is a measure of how negative, positive, or neutral any tokenized part of some textual data is. Polarity helps convey the overall emotional weights or importance that any particular sentence might have to the sentiment analyzer [10].

Polarity by itself is not an absolute measure of the overall sentiment associated with any sentence, for example: words like 'superb' or 'good' might indicate the sentence has a positive sentiment, words like 'bad' or 'sad' might indicate a negative sentiment, but what would happen in case the sentence has a combination of them, like 'not bad' will we classify it as neutral or opposite of bad ?

Polarity is measured on a normalized scale of -1 to +1. The Sentiment Analyzer API then measures, combines, and normalizes values on both the polarity of the overall text, individual sentences, and individual phrases. This returns a better picture of the relative polarities of texts by not penalizing longer sentences that are expressing positive or negative emotion at scale but also contain neutral phrases. The Polarity scores for our dataset are shown below in figure 8.

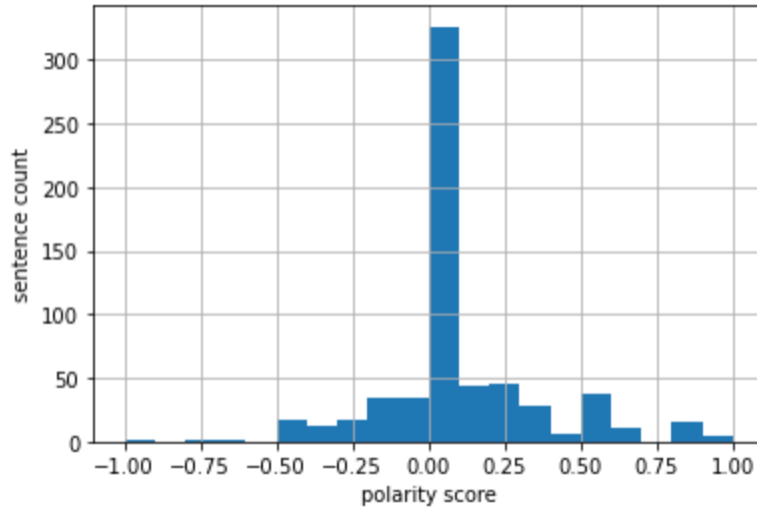


Figure 8: Polarity Histogram plot of our Gathered Data

Subjectivity Score: Text subjectivity is a measure of how subjective or objective the statement is. An objective statement has presumably true factual information. A subjective statement gives an opinion about something, that opinion may or may not be factually true. This is much harder to model [11].

Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. The image below (figure 9) shows subjectivity scores for our dataset.

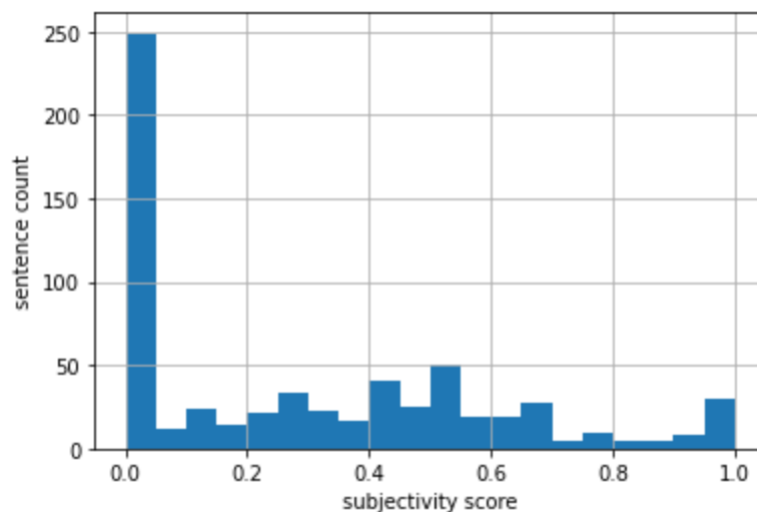


Figure 9: Subjectivity Histogram plot of our Gathered Data

Data Cleaning and Transformation

Data Collection :- We use a Twitter developer account to download recent tweet data and to achieve this we utilize the Tweepy library APIs (Application programming interface). Our data is gathered in real-time, and it takes roughly 7-9 minutes each run because of data read speed restrictions imposed by Twitter.

```
# This block will authorize twitter API and connect it to my developer account

access_token = "1574192041273094144-9qUjidfpWYpLyLxMAo4b4BMmt401xa"
access_token_secret = "uLxLAUxNpTy02PFq8ZXv1jK16GRDuUZEPLCN0rMd8Y4Uu"
consumer_key = "bvQ8gbYfpUvZ8Y0v4dWCUwBLj"
consumer_secret = "reMeIYZKAlyz9wVRo8XDW1xN6t0w14C0wkM9rtBZMSz58JpWPn"

auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tw.API(auth)
```

Figure 10: Python code to authorize tweepy API

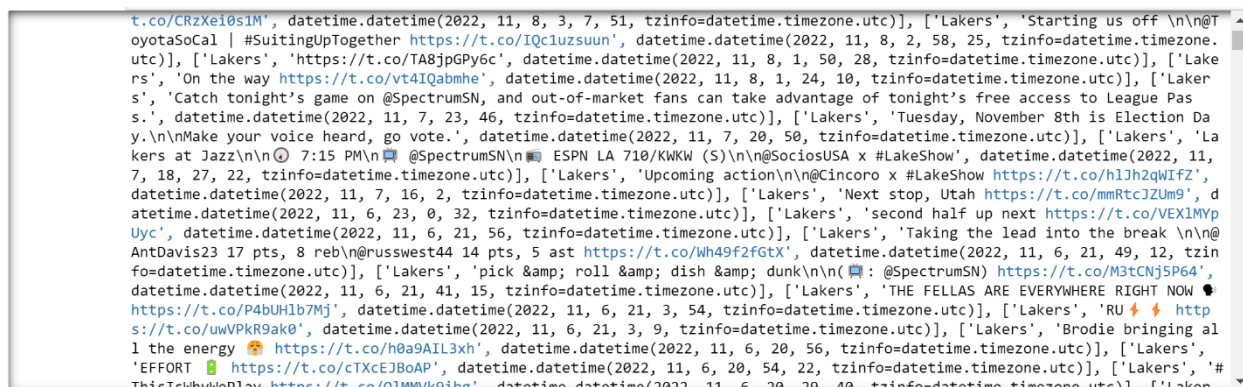
```
# This block will collect tweets and store them in gathered_tweets list.

gathered_tweets = []
# It will take 1000 tweets from each handle
for u in users:
    tweets_user_1 = tw.Cursor(api.user_timeline,screen_name=u).items(1000)
    time.sleep(3)
    x = [[tweet.user.screen_name, tweet.text, tweet.created_at] for tweet in tweets_user_1]
    gathered_tweets.append(x)
    time.sleep(3)
```

Figure 11: Using Tweepy to parse tweets online and store them

Since this is real world data this will not be clean, there will be a lot of emoticons, special characters, numbers, language queues, and more inside these tweets. In order to be able to make useful inferences with the gathered information we need to remove these special characters and transform our dataset.

Data Transformation :- When using the Tweepy API, we store our dataset in the form of a list. The raw gathered data is shown below.



```
t.co/CR2Xe10s1M', datetime.datetime(2022, 11, 8, 3, 7, 51, tzinfo=datetime.timezone.utc)), ['Lakers', 'Starting us off \n\n@T
oyotaSoCal | #SuiingUpTogether https://t.co/IQc1uzsuun', datetime.datetime(2022, 11, 8, 2, 58, 25, tzinfo=datetime.timezone.
utc)), ['Lakers', 'https://t.co/TA8jpGPy6c', datetime.datetime(2022, 11, 8, 1, 50, 28, tzinfo=datetime.timezone.utc)), ['Laker
rs', 'On the way https://t.co/vt4IQabmhe', datetime.datetime(2022, 11, 8, 1, 24, 10, tzinfo=datetime.timezone.utc)), ['Laker
s', 'Catch tonight's game on @SpectrumSN, and out-of-market fans can take advantage of tonight's free access to League Pas
s.', datetime.datetime(2022, 11, 7, 23, 46, tzinfo=datetime.timezone.utc)), ['Lakers', 'Tuesday, November 8th is Election Da
y.\n\nMake your voice heard, go vote.', datetime.datetime(2022, 11, 7, 20, 50, tzinfo=datetime.timezone.utc)), ['Lakers', 'La
kers at Jazz\n\n 7:15 PM\n @SpectrumSN\n ESPN LA 710/KWKK (S)\n\n@SociosUSA x #LakeShow', datetime.datetime(2022, 11,
7, 18, 27, 22, tzinfo=datetime.timezone.utc)), ['Lakers', 'Upcoming action\n\n@Cincoro x #LakeShow https://t.co/h13h2qWfZ',
datetime.datetime(2022, 11, 7, 16, 2, tzinfo=datetime.timezone.utc)), ['Lakers', 'Next stop, Utah https://t.co/mmRtcJZUm9', d
atetime.datetime(2022, 11, 6, 23, 0, 32, tzinfo=datetime.timezone.utc)), ['Lakers', 'second half up next https://t.co/VEX1Myp
Uyc', datetime.datetime(2022, 11, 6, 21, 56, tzinfo=datetime.timezone.utc)), ['Lakers', 'Taking the lead into the break \n\n@
AntDavis23 17 pts, 8 reb\n@russwest44 14 pts, 5 ast https://t.co/Wh49f2fGTX', datetime.datetime(2022, 11, 6, 21, 49, 12, tzin
fo=datetime.timezone.utc)), ['Lakers', 'pick & roll & dish & dunk\n\n( @SpectrumSN) https://t.co/M3tCNj5P64',
datetime.datetime(2022, 11, 6, 21, 41, 15, tzinfo=datetime.timezone.utc)), ['Lakers', 'THE FELLAS ARE EVERYWHERE RIGHT NOW
https://t.co/P4bUH1b7Mj', datetime.datetime(2022, 11, 6, 21, 3, 54, tzinfo=datetime.timezone.utc)), ['Lakers', 'RU
https://t.co/uwVPkR9ak0', datetime.datetime(2022, 11, 6, 21, 3, 9, tzinfo=datetime.timezone.utc)), ['Lakers', 'Brodie bringing al
l the energy https://t.co/h0a9AIL3xh', datetime.datetime(2022, 11, 6, 20, 56, tzinfo=datetime.timezone.utc)), ['Lakers',
'EFFORT https://t.co/CTXcE3BoAP', datetime.datetime(2022, 11, 6, 20, 54, 22, tzinfo=datetime.timezone.utc)), ['Lakers', '#
ThisIsWhvWePlav https://t.co/Q1MMVv9ibe', datetime.datetime(2022, 11, 6, 20, 29, 40, tzinfo=datetime.timezone.utc)), ['Laker
```

Figure 12: Raw Gathered Data

We utilize the Pandas library to convert this data into an organized dataset. The Tweets list transformed into a dataset is shown in Figure 13.

	User	Tweet	Date
0	Lakers	Brodie ballin' off the bench	2022-11-15 21:37:22+00:00
1	Lakers	RT @mitchell_ness: The @Lakers' Superman\n\nIn...	2022-11-15 21:20:22+00:00
2	Lakers	2017-18 City Edition vs. Showtime Purple	2022-11-15 20:26:21+00:00
3	Lakers	Time to decide the GOAT Lakers jersey \n\n@goa...	2022-11-15 20:26:20+00:00
4	Lakers	"He's been a monster." \n\n@AntDavis23 with th...	2022-11-15 01:04:06+00:00
...
10685	LakersCommunity	RT @LakerGirls: A special morning with the US ...	2015-11-11 19:19:31+00:00
10686	LakersCommunity	RT @LosLakers: ¡Feliz día de los veteranos! #...	2015-11-11 19:19:25+00:00
10687	LakersCommunity	RT @DFenders: Thanks to the US Vets of Inglewo...	2015-11-11 18:21:10+00:00
10688	LakersCommunity	RT @LakerGirls: Honored to be spending this Ve...	2015-11-11 17:22:06+00:00
10689	LakersCommunity	.@lakers staffers are out volunteering at the ...	2015-11-11 17:21:59+00:00

10690 rows × 3 columns

Figure 13: Organized Dataset

Feature Selection

Data Cleaning :- At this point this data is still not useful for the sentiment analyzer and so we must remove all the special characters and prepare this dataset better, we start by removing a few special groups of characters outlined in figure 14 below.

```
def remove(x):
    emoji_pattern = re.compile("[
        u\"\\U0001F600-\\U0001F64F\" # emoticons
        u\"\\U0001F300-\\U0001F5FF\" # symbols & pictographs
        u\"\\U0001F680-\\U0001F6FF\" # transport & map symbols
        u\"\\U0001F1E0-\\U0001F1FF\" # flags (iOS)
        u\"\\U00002500-\\U00002BEF\" # chinese char
        u\"\\U00002702-\\U000027B0\" # flags (Android)
        u\"\\U000024C2-\\U0001F251\" # positional/gps symbols
        u\"\\U0001F926-\\U0001F937\" # Emoji set 2
        u\"\\U00010000-\\U0010ffff\" # Emoji set 3
        u\"\\u2640-\\u2642\" # Emoji set 4
        u\"\\u2600-\\u2B55\" # Emoji set 5
        u\"\\u200d\" # Zero width joiner
        u\"\\u23cf\" # Eject symbol
        u\"\\u23e9\" # black right-pointing double triangles
        u\"\\u231a\" # Unicode escape
        u\"\\ufe0f\" # dingbats
        u\"\\u3030\" # wavy dash
    ]+", re.UNICODE)
    x = re.sub(emoji_pattern, '', x)
    x = re.sub(r'http\S+', '', x)
    return x

tweet_data['Tweet'] = tweet_data['Tweet'].apply(remove)
# print(tweet_data.head(10))
```

Figure 14: Python function for removing special characters.

One thing to note is that there are some special characters that are ignored when cleaning our data at this stage, these are new line characters like “\n\n” or usernames associated with the Tweets, or the “@” symbol used to designate twitter handles as usernames. In later stages when remove the punctuations and stopwords specific to English language that is when we can remove the aforementioned special characters.

Calibrating Data for Recent Events :- The use case for our project is to evaluate the recent trends regarding any subject of choice, for instance we chose the LA Lakers. And so to achieve this we restrict our data gathering to only the recent events that have happened relating our subject withing the last 1 week from running our data analysis model. The python code we used to achieve this is shown below in figure 15.

```
# making sure only 1 week of data is parsed from dataframe with clean data

start_date = date.today()
days = timedelta(days = 7)
endDate = start_date - days

tweet_data['Date'] = pd.to_datetime(tweet_data['Date']).dt.date

tweet_data = tweet_data[ (tweet_data['Date'] < start_date)
                        & (tweet_data['Date'] > endDate)]
```

Figure 15: Python code to remove old data.

The Dataset after this operation is as shown below in figure 16.

	User	Tweet	Date
0	Lakers	Brodie ballin' off the bench	2022-11-15
1	Lakers	RT @mitchell_ness: The @Lakers' Superman\n\nIn...	2022-11-15
2	Lakers	2017-18 City Edition vs. Showtime Purple	2022-11-15
3	Lakers	Time to decide the GOAT Lakers jersey \n\n@goa...	2022-11-15
4	Lakers	"He's been a monster." \n\n@AntDavis23 with th...	2022-11-15
...
8742	thelakers248	Lakers News: The Only Way LeBron James Thinks ...	2022-11-10
8743	thelakers248	Lakers News: LeBron James Isn't Just Consideri...	2022-11-10
8744	thelakers248	Lakers News: Anthony Davis Does Not Mince Word...	2022-11-10
9690	LakersCommunity	A night of Thanksgiving with the Lakers! \n\nl...	2022-11-14
9691	LakersCommunity	Happy Veteran's Day \n\nWe are wrapping up #H...	2022-11-12

All Dates within 1 week



Figure 16: Dataset after being restricted to 1 week.

Discretizing Data :- We are using the NLTK library for performing the sentiment analysis task, and for the NLTK analyzer to work efficiently we need to now tokenize all the different words inside each sentence in our dataset. To achieve this we use

the 'word_tokenize' function from the NLTK library. The python code to achieve this task is shown below in figure 17.

```
words = nltk.word_tokenize(str(tweet_data).lower())
```

Figure 17: Tokenizing our dataset

Tokenizing our dataset in this manner helps us to achieve two main tasks:

- It helps us in normalizing the dataset later when we remove any punctuation and stopwords specific to the English language.
- It helps us in the feature extraction process which involves lemmatization, stemming and the subjectivity and polarity scores.

Stop Word Removal :- Stop words are common words like 'the', 'and', 'is', 'for' etc. that are very frequent in text, and so they do not present or offer any significant insight into the specific topic of the document. We must remove these stop words from the text in a given corpus to clean up the data, and easily identify words that are more rare and potentially more relevant to what we're interested in.

The process for stop word removal starts by first tokenizing the words by using the nltk word_tokenize function. The list of stop words for removal is sourced from the NLTK English library. But we also add a few stop words from our side as they were appearing much more frequently than other words, they were "https" and "lakers". Before ending the stop word removal stage we also manually remove all punctuations and digits from our dataset by using an iterative for loop and empty lists.

The frequency plots of the 20 most frequent words before and after the stop word removal are shown below in figure 18.

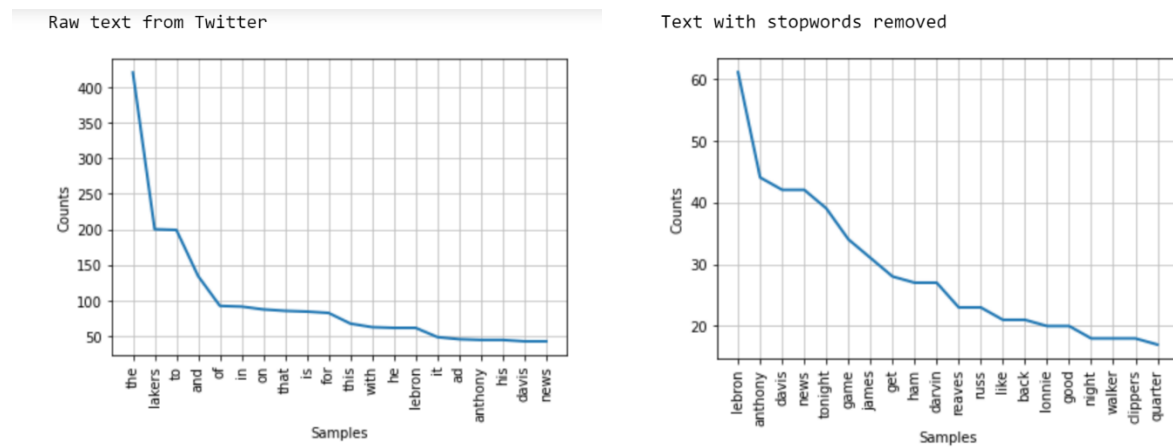


Figure 18: Raw text(left) and Data with stopwords removed (right).

Feature Extraction

Lemmatization :- Lemmatisation (or lemmatization) in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.

In computational linguistics, lemmatisation is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatisation depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document.

For our case the specific rules by which the lemmatization functions work is stored inside the NLTK library, they are a part of the same English language corpus we used

for stopwords removal. The figure below depicts the theoretical outcome expected from a lemmatization function and the results that we got.

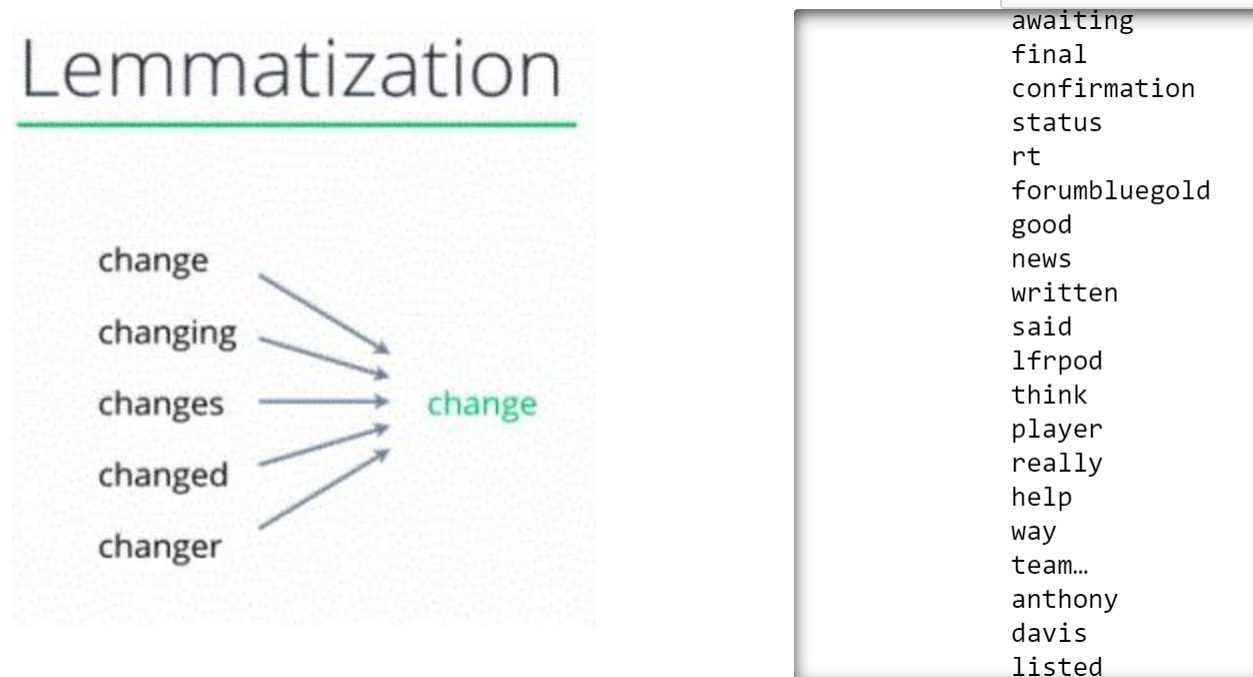


Figure 19: Lemmatization theoretical (left) and actual (right).

Stemming :- In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation. Figure 20 shows the theoretical concept of stemming in a more detailed manner, and we can also see the outcome we achieved after this process.

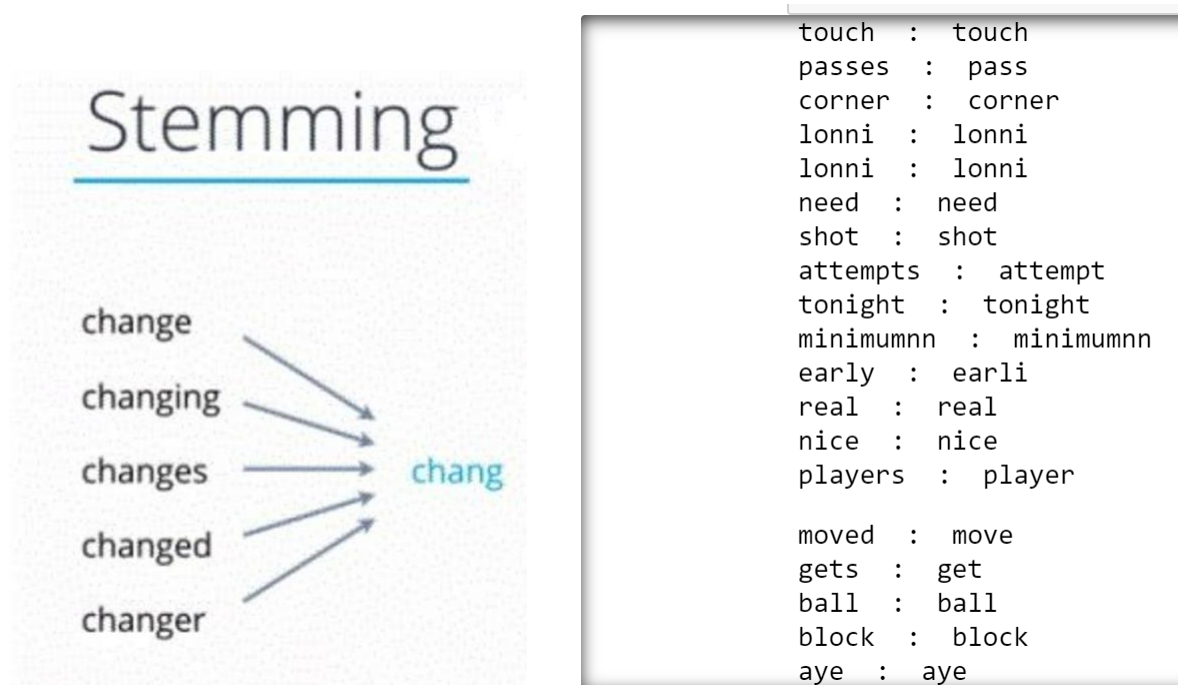


Figure 20: Stemming theoretical (left) and actual (right).

Polarity :- The definition of Polarity has been explained in an earlier section in this report, here we will look at how we can use polarity to enhance our feature extraction efforts. Here we can also get a pre-emptive indication of whether our efforts with the data preprocessing have been useful or not, by comparing the polarity score of our dataset before and after all the previously discussed data preprocessing techniques. Figure 21 shows the polarity score of the original raw data on the left and the processed data on the right.

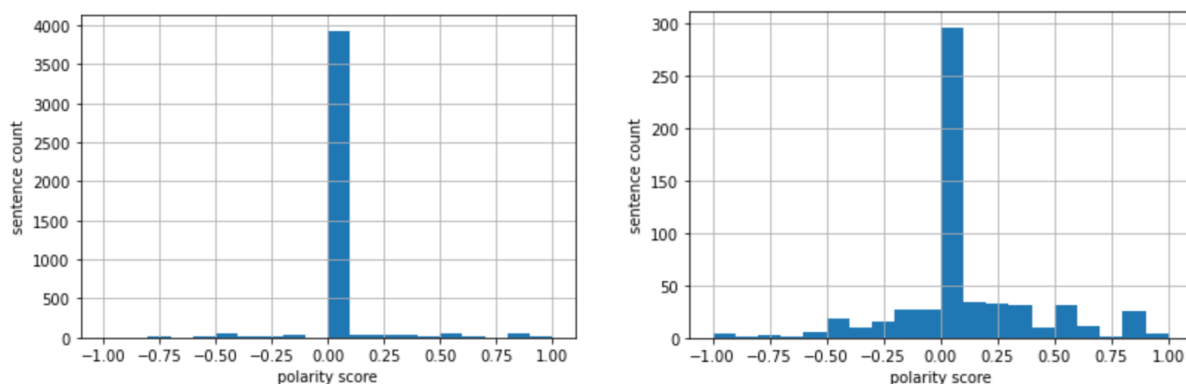


Figure 21: Polarity plot of original dataset (left) and processed dataset (right).

Subjectivity :- The definition of Subjectivity has been explained in an earlier section, here we will look at how we can use Subjectivity to enhance our feature extraction efforts. Similar to how we assess data pre-processing using polarity we can do so with subjectivity as well, the figure below shows the subjectivity scores for our dataset before (on the left) and after data pre-processing (on the right).

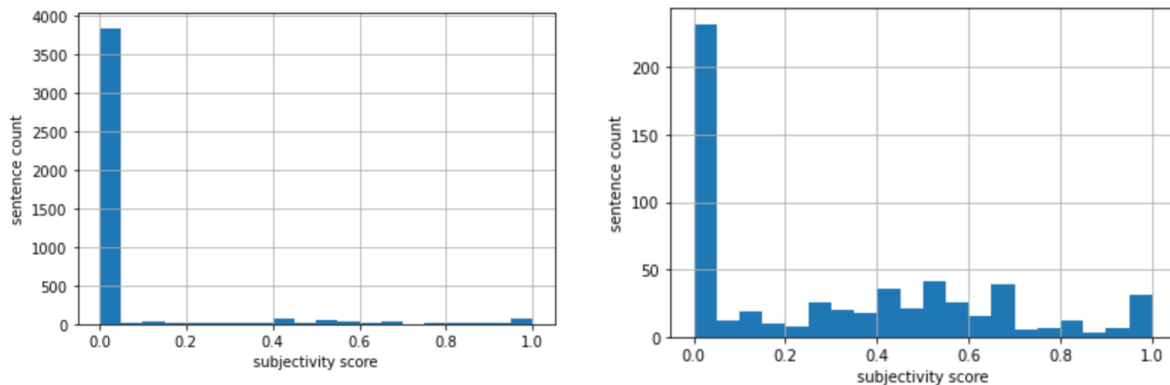


Figure 22: Subjectivity plot of original dataset (left) and processed dataset (right).

Modeling

Text communication is one of the most popular forms of day to day conversation. We chat, message, tweet, share status, email, write blogs, share opinion and feedback in our daily routine. All of these activities are generating text in a significant amount, which is unstructured in nature. In this area of the online marketplace and social media, It is essential to analyze vast quantities of human generated textual data, to understand peoples opinion.

For the modelling aspect of our Sentiment Analysis Project we have chosen to utilize the “Natural Language Toolkit” Python Library, more commonly known as the NLTK library. More details regarding the NLTK library have been discussed below.

NLTK :- The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. It was developed by Steven

Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania. NLTK includes graphical demonstrations and sample data. It is accompanied by a set of rules and stipulations regarding the English language that is intended to guide the NLTK functions in making useful inferences, this set of rules and stipulations is often referred to as a Lexicon.

NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems. There are 32 universities in the US and 25 countries using NLTK in their courses. NLTK supports classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities.

NLTK is a powerful Python package that provides a set of diverse natural languages algorithms. It is free, opensource, easy to use, large community, and well documented. NLTK consists of the most common algorithms such as tokenizing, part-of-speech tagging, stemming, sentiment analysis, topic segmentation, and named entity recognition. NLTK helps the computer to analysis, preprocess, and understand the written text.

Evaluation

Evaluating the performance of our sentiment analysis modelling techniques the path is not so straight forward, this is because we are dealing with uncategorized data which is in the form of textual tweets made by users on the Twitter platform and this data is collected in real-time whenever we run the model to perform the sentiment analysis. Therefore these tweets when parsed from the Twitter database will not come with sentiments already associated to these Tweets. To elaborate this point: Twitter is most often seen as a free-speech platform and so Twitter does not associate positive, negative or neutral connotations with tweets made by its users as that would be against its policy, hence any Tweet that you see online is not associated with any kind of positive, negative or neutral sentiment tag. And since the sentiment associated with the tweets is our final goal this means we cannot

rely on Supervised Machine Learning approaches meaning the majority of conventional evaluation approaches get thrown out the window and we must instead think of other possible alternatives.

Two major ideologies that we have used here will be categorized into a direct and indirect approach. The direct approach will involve using the original dataset that we collect in real-time from twitter, this will involve utilizing the Polarity and Subjectivity scores which are a part of the NLTK library functions. The indirect approach will involve the use of an external labelled dataset which is sourced from Github and has already labelled target values in the form of 'positive', 'negative' and 'neutral' values associated with each tweet, we can then remove and separately store the target values before passing this external dataset through our sentiment analysis model and based on the generated target values we can create a 3x3 Multi Class Confusion Matrix to evaluate conventional performance measures like Accuracy, Precision, Recall and the F1 score.

Direct Approach

This approach involves using the dataset gathered in real-time from Twitter and utilizing the Polarity score and Subjectivity scores as performance measures. Though these are not ideal performance measures in the sense that we don't get an empirical values like we do in the case of accuracy, recall and other conventional performance measures. But we do get a sense of how much increase in optimization or performance improvement we get by comparing the before and after of data pre-processing functions. This can give us a sense of how much our modelling has improved because of the functions performed in the data pre-processing part of the project.

Polarity :- Polarity can also act as an evaluation tool for our NLTK model. Polarity is at its core a measure of how positive, negative or neutral any individual word or sentence is in a dataset, this can be extended to apply to individual Tweets in our dataset and thus plays a vital role in signaling the sentiment analyzer whether a sentence/tweet as a whole should be categorized as positive, negative or neutral. It can also be used as an indirect indicator of our model's performance as we have done here. Figure 23 shows the polarity score of the original raw data on the left and the processed data on the right.

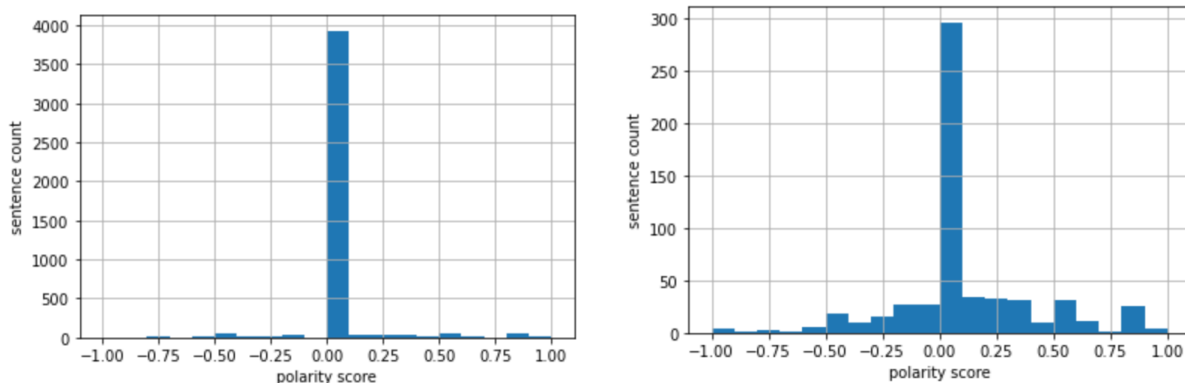


Figure 23: Polarity plot of original dataset (left) and processed dataset (right).

Subjectivity :- Text subjectivity is a measure of how subjective or objective the statement is. An objective statement has presumably true factual information. A subjective statement gives an opinion about something, that opinion may or may not be factually true. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. Similar to how we asses data pre-processing using polarity we can do so with subjectivity as well, the figure below shows the subjectivity scores for our dataset before (on the left) and after data pre-processing (on the right).

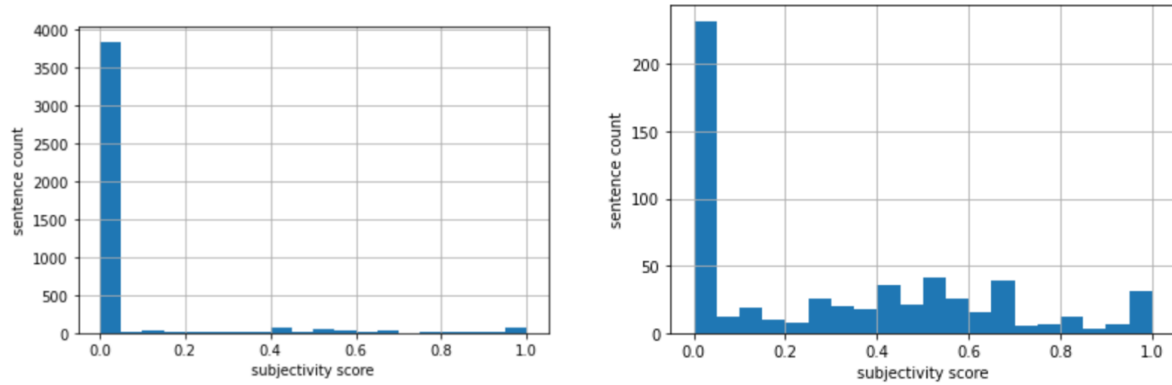


Figure 24: Subjectivity plot of original dataset (left) and processed dataset (right).

Indirect Approach

This approach involves using an external dataset, one which is sourced from Github and the link to which is added [here](#). This dataset is one which already has sentiments associated to them in the form of a target feature added to the analyzed tweets, this is akin to having a labelled dataset which in turn allows us to utilize conventional performance measures like accuracy, precision, recall and the F1 score. A small part of the dataset used here has been shown in the image below.

329	0	2177	Mon May 25 17 18 49 UTC 2009	north korea	one_eighteen	Oooooooh... North Korea is in troubleeeee	19ep4H
330	0	2178	Mon May 25 17 19 07 UTC 2009	north korea	FOLKTALE09	Wat the heck is North Korea doing	They just conducted powerful nuclear tests Follow the li
331	0	2179	Mon May 25 17 19 30 UTC 2009	north korea	Mvsic	Listening to Obama... Friggin North Korea...	
332	0	2180	Mon May 25 17 21 16 UTC 2009	pelosi	CFURNAROS	I just realized we three monkeys in the white Obama.Biden,Pelosi . Sarah Palin 2012	
333	0	2181	Mon May 25 17 21 30 UTC 2009	pelosi	Rachael90210	foxnews Pelosi should stay in China and never come back.	
334	0	2182	Mon May 25 17 21 35 UTC 2009	pelosi	TylerSchmidt	Nancy Pelosi gave the worst commencement speech I ve ever heard. Yes I m still bitter about this	
335	0	2183	Mon May 25 17 25 36 UTC 2009	insects	KayJay_x	ugh, the amount of times these stupid insects have bitten me. Grr..	
336	4	2184	Mon May 25 17 25 54 UTC 2009	insects	BecCrew	Prettiest insects EVER - Pink Katydids	2Upw2p
337	0	2185	Mon May 25 17 26 30 UTC 2009	insects	euthanasia86	Just got barraged by a horde of insects hungry for my kitchen light. So scary.	
338	4	2187	Mon May 25 17 29 06 UTC 2009	mcdonalds	connlocks	Just had McDonalds for dinner. D It was gooooood. Big Mac Meal.	
339	4	2188	Mon May 25 17 29 11 UTC 2009	mcdonalds	MamiYessi	AHH YES LOL IMA TELL MY HUBBY TO GO GET ME SUM MCDONALDS	
340	4	2190	Mon May 25 17 29 46 UTC 2009	mcdonalds	Yuleineeee	Stopped to have lunch at McDonalds. Chicken Nuggetssss yummmmy.	
341	4	2191	Mon May 25 17 29 51 UTC 2009	mcdonalds	XrachuIX	Could go for a lot of McDonalds. i mean A LOT.	
342	4	2193	Mon May 25 17 31 52 UTC 2009	exam	xKimmellie	my exam went good. HelloLeonie your prayers worked	
343	4	2194	Mon May 25 17 31 58 UTC 2009	exam	laulaulauren	Only one exam left, and i am so happy for it D	
344	0	2195	Mon May 25 17 32 22 UTC 2009	exam	elllllen	Math review. Im going to fail the exam.	
345	0	2196	Mon May 25 17 35 08 UTC 2009	cheney	LPSsports43	Colin Powell rocked yesterday on CBS. Cheney needs to shut the hell up and go home.Powell is a man of H	
346	0	2197	Mon May 25 17 35 43 UTC 2009	cheney	joahs	obviously not siding with Cheney here	19j2d
347	4	2202	Tue May 26 22 39 46 UTC 2009	mashable	paulobsf	Absolutely hilarious	from mashable bccwt
348	4	2203	Tue May 26 22 40 41 UTC 2009	mashable	christinerose	mashable I never did thank you for including me in your Top 100 Twitter Authors	You Rock I
349	2	2204	Wed May 27 00 34 45 UTC 2009	jquery book	cfbloggers	Learning jQuery 1.3 Book Review -	cfbloggers.org c30629
350	4	2205	Wed May 27 00 37 30 UTC 2009	jquery book	pdelsignore	RT shrop Awesome JQuery reference book for Coda	www.macpeeps.com coda webdesign
351	4	2206	Wed May 27 00 38 44 UTC 2009	goodby silverstein	bskatz	I ve been sending e-mails like crazy today to my contacts...does anyone have a contact at Good	

Figure 25: Segment of the external dataset used for evaluation purposes.

A few things to note about this external dataset are :-

- It has sentiment labels associated with tweets, the orange column in figure 3 shows the associated sentiments, 4 refers to a positive sentiment, 2 refers to a neutral sentiment and 0 refers to a negative sentiment.

- This dataset is close to an equal distribution for each of the target sentiment of positive, negative and neutral making it an ideal dataset for evaluation purposes, more specifically this dataset has 498 tweets in total, 182 of which are positive sentiments, 177 negative sentiments and 139 neutral sentiments.
- The red box highlights some of the negative sentiment tweets, the blue box highlights some of the positive sentiment tweets and the green box shows what a neutral sentiment tweet looks like.

We mentioned using conventional evaluation approaches but one thing we need to keep in mind is that since our target value (sentiment) can have one of three possible outcomes we cannot use the typical 2x2 Confusion Matrix, instead we rely on a 3x3 Multi-Class Confusion Matrix for our purposes.

3x3 Confusion Matrix :- Before we look at the confusion matrix, we obtained using the new dataset and our own sentiment analyzer model that uses NLTK, we will look at how we draw performance inferences from a 3x3 Confusion Matrix. Figure 26 below shows a typical 3x3 confusion matrix (left) and it also shows us how we obtain the Recall and Precision scores using that matrix (right).

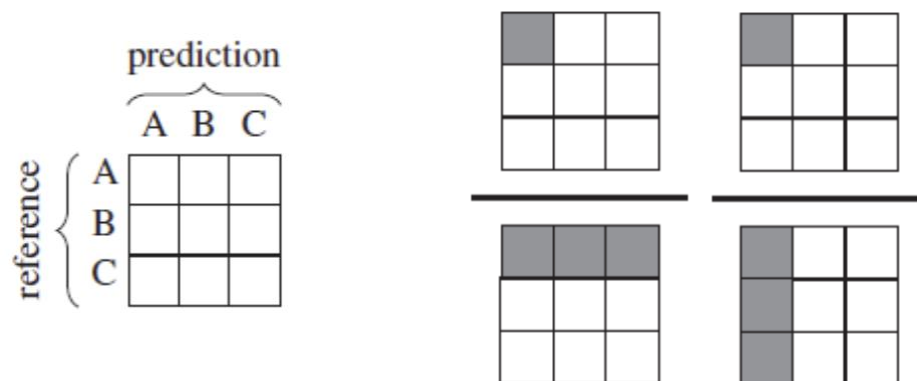


Figure 26: 3x3 Confusion Matrix (left), Recall (middle) and Precision Score (right).

Unlike the traditional 2x2 Confusion Matrix the precision and recall calculations are slightly different here, as shown in the middle part of figure 4, to calculate the recall score we will divide the number present in position AA of the matrix with the sum of the elements in row A. Similarly for calculating the precision as shown in the leftmost part of figure 27, we will divide the number in position AA with the sum of all elements in the column A.

Now we can look at the 3x3 confusion matrix we obtained using our NLTK model and draw performance evaluations for the model. Figure 5 below shows the confusion matrix we obtained after removing the sentiment data from the external dataset and doing the sentiment analysis ourselves using our NLTK model.

CONFUSION MATRIX : TEST DATASET

TARGET LABEL	pos	168	1	5
	neu	10	135	12
	neg	4	3	160
		pos	neu	neg

PREDICTION LABEL

Figure 27: 3x3 Confusion Matrix for external dataset using our NLTK model.

Using this 3x3 Confusion Matrix we make the following inferences :-

- | | | |
|--|--|---|
| <ul style="list-style-type: none"> ▪ Precision_{positive} = 0.92 ▪ Precision_{neutral} = 0.97 ▪ Precision_{negative} = 0.87 ▪ F₁ (positive) = 0.93 ▪ F₁ (neutral) = 0.90 ▪ F₁ (negative) = 0.90 ▪ F₁ (average) = 0.91 ▪ Accuracy (overall) = 0.929 | | <ul style="list-style-type: none"> Recall_{positive} = 0.96 Recall_{neutral} = 0.85 Recall_{negative} = 0.95 |
|--|--|---|

These results showcased above lead us to believe that our NLTK sentiment analysis model performs adequately.

Deployment

Since Twitter acts as a repository of all old tweets made by users and they are stored there for posterity unless the user wishes to go back and delete the tweet they made, this allows us to calibrate our model and allows it to look at past events related to LA Lakers, this can serve as yet another pseudo performance check if the results of our sentiment analysis can correlate to past events surrounding the LA Lakers team.

Our model only analyzes sentiments that range within 1 week and relate to the LA Lakers Basketball team. Figure 28 shows the sentiment analysis results correlated with actual events that took place within the LA Lakers team.

{ 'neg' : 0.102, 'neu' : 0.465, 'pos' : 0.587, 'compound' : 0.721 }

Analysis Week : 3rd week of June
Lakers had just finished the NBA Summer League with 3 wins
and 2 losses in the season. Predicted public opinion is
somewhere between positive and neutral.

{ 'neg' : 0.682, 'neu' : 0.421, 'pos' : 0.276, 'compound' : 0.238 }

Analysis Week : 2nd week of October
halfway through the 2nd week of October the preseason
matches have ended, Lakers had a poor performance record
with only 1 win out of the 6 matches that they played.
Predicted public opinion is tending towards the negative side.

{ 'neg' : 0.302, 'neu' : 0.365, 'pos' : 0.627, 'compound' : 0.866 }

Analysis Week : 1st week of December
The season so far for the Lakers has not gone well, they have
lost 11 out of the 16 matches they played up until 22nd
November, but in the next 6 matches they have managed to
win 5 of them. Predicted public opinion is swaying back
towards the positive side.

Figure 28: Deploying our model across different weeks of 2022 for analysis.

Summary of Learning Experiences

During this group project we learned a great deal together, some of the most important things and most crucial lessons we learned are listed below.

- Breaking complex tasks into parts and steps so that we can achieve them easily.
- Planning and Managing time so that we can work on the tasks together and so that the burden on any 1 person is not overwhelming.
- Refining understanding through discussions and explanations so that even if any member of the group is not familiar with the topic of discussion he/she can get a good understanding and contribute in their own way.
- Developing stronger communication skills, as all of us come from different academic and social backgrounds and hence by working together on a common project we developed better communication skills.
- We learned the importance of delegating roles and responsibilities so that we may achieve our goals in an orderly and timely manner.
- Pooling knowledge and skills helped us overcome the major hurdles we faced in our project, having varied knowledge from different fields it was important for us to combine efforts so that where one person might be weak, the other person can take things forward in a more meaningful manner.

References

- [1] S. Bhuta, A. Doshi, U. Doshi and M. Narvekar, "A review of techniques for sentiment analysis Of Twitter data," 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014, pp. 583-591, doi: 10.1109/ICICT.2014.6781346.
- [2] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," 2014 Seventh International Conference on Contemporary Computing (IC3), 2014, pp. 437-442, doi: 10.1109/IC3.2014.6897213.
- [3] <https://www.questionpro.com/blog/twitter-sentiment-analysis/#:~:text=A%20Twitter%20sentiment%20analysis%20is,attitudes%2C%20emotions%2C%20and%20opinions.>
- [4] <https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/>
- [5] <https://monkeylearn.com/blog/sentiment-analysis-of-twitter/>
- [6] <https://huggingface.co/blog/sentiment-analysis-twitter>
- [7] <https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf>
- [8] <https://developer.twitter.com/en/docs/tutorials/how-to-analyze-the-sentiment-of-your-own-tweets>
- [9] <https://ojs.aaai.org/index.php/ICWSM/article/view/14185>
- [10] M. Desai and M. A. Mehta, "Techniques for sentiment analysis of Twitter data: A comprehensive survey," 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016, pp. 149-154, doi: 10.1109/CCAA.2016.7813707.
- [11] V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPEs), 2016, pp. 1345-1350, doi: 10.1109/SCOPEs.2016.7955659.

Appendix

The Source code for this project has been uploaded to my Github repository, the Github link is:

https://github.com/hardiksingh933/Analyzing_Twitter_Population_Sentiments_based_on_LA_Lakers

Please note that you will not find any separate folder containing the dataset in the Github repository as our project includes collecting and organizing the dataset on our own from Twitter servers, the initial part of our Python code which involves dealing with the Tweepy library and the authentication keys that are linked to my personal Twitter Developer Account are responsible for collecting the Raw Data from Twitter, furthermore our Python source code also deals with the cleaning and proper formatting of the dataset as explained in earlier sections.