Twitter Data Analysis

CS 226: BIG-DATA MANAGEMENT, Fall 2022

Group number 11: Fantastic4

Vineet Dhaimodker

Vishv Patel

Mihir Patel

Yash Gandhi

FIFA WORLD CUP Qatar2022

Table of Contents

Background & Motivation

Relevance of our work

Overview of work done

Data Gathering

Data Pre-processing

Sentiment analysis

Visualization Dashboard Demo

Evaluation

Related Work

Conclusion

FIFA WORLD CUP Qatar2022

Background & Motivation

1

- FIFA World Cup is the most viewed event with about **3.5 billion** followers worldwide and the 2018 FIFA World Cup generated more than **5 billion** USD of revenue in total
- Twitter conversations regarding the same keep growing with a total of **41 million** tweets related to soccer in UK alone, since the beginning of the year

2

- Visualizing twitter data can benefit the marketers to analyze reports, interests, evaluate performance of strategies
- Sentiment analysis helps to find hidden patterns such as brand perception which is a massive factor for large businesses

Relevance of work

- ➔ Perfect example of big data as it follows the four V's of Big Data
- ➔ Excellent source to understand the ground truths behind occurrence of global events
- ➔ Businesses use it for marketing, growth and development and take feedback from their followers for consistent improvement

Relevance of work

- ➔ Perfect example of big data as it follows the four V's of Big Data
- ➔ Excellent source to understand the ground truths behind occurrence of global events
- ➔ Businesses use it for marketing, growth and development and take feedback from their followers for consistent improvement

Data size

~14GB

Total size of tweet data gathered

Relevance of work

- ➔ Perfect example of big data as it follows the four V's of Big Data
- ➔ Excellent source to understand the ground truths behind occurrence of global events
- ➔ Businesses use it for marketing, growth and development and take feedback from their followers for consistent improvement

Soccer related tweets

5M

FIFA World cup related Tweets gathered since the beginning of 2022

Data size

~14GB

Total size of tweet data gathered

Relevance of work

- ➔ Perfect example of big data as it follows the four V's of Big Data
- ➔ Excellent source to understand the ground truths behind occurrence of global events
- ➔ Businesses use it for marketing, growth and development and take feedback from their followers for consistent improvement

Soccer related tweets

5M

FIFA World cup related Tweets gathered since the beginning of 2022

Data size

~14GB

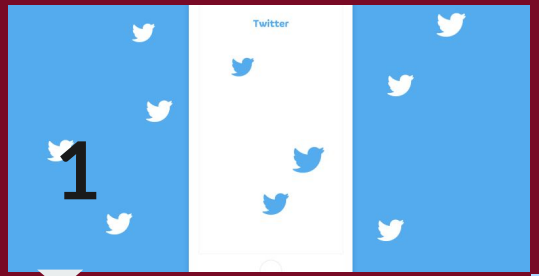
Total size of tweet data gathered

Relevant hashtags

53

Trending twitter hashtags for FIFA Worldcup 2022

Project Overview

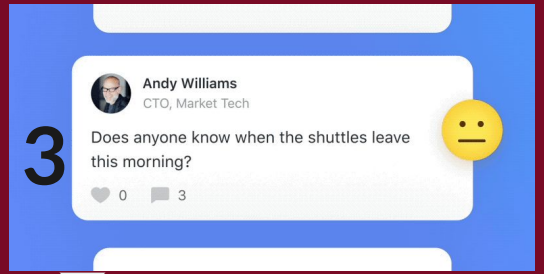


Data Gathering

Extracted FIFA 2022 World cup tweets for trending hashtags using snscrape. Collected almost 14GB data

Data Preprocessing

PySpark and Hadoop for handling and processing the big data. NLTK libraries to process tweets for Sentiment Analysis.



Sentiment Analysis

We will be using Machine learning model on Tweets to classify the Sentiments such as Logistic Regressio and Naive Bayes.

Project Overview

4

Evaluation

We evaluate the sentiment analysis models using standard evaluation metrics. We also calculate the query response time to evaluate scalability

6

Data Visualization

React JS with its map and chart components helps to create a dashboard to visualize distribution of tweets based on sentiments, topics, mapping tweets on world map

5

Conclusion

Supervised ML models were used to learn sentiments from the custom dataset and the accuracy of Naïve Bayes and Logistic regression was evaluated and compared

Data Gathering - SNScrape

- Disadvantage of Twitter API
  - limitations on the number of requests : 900 requests/15-minutes
- Timeline of data extraction: January 2022 – November 2022
- Hashtags
  - Trending: fifaworldcup, FIFAWorldCup2022, FIFAWorldCupQatar2022
  - Teams: threelions (England), usmnt (USA)
  - Matches: qatarvsecuador, qtrecu, porarg
- Extracted ~14 GB of twitter data with almost 5 million tweets.

Data Pre-processing: technology stack

1. We store the data in the Hadoop cluster. This replicates the data into multiple datanodes.

2. We used spark to process the data which we take from the HDFS and used PySpark for further processing the data

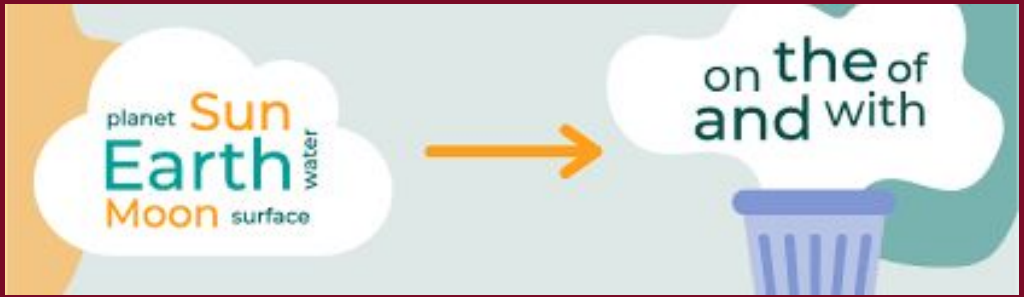
Data Pre-processing

1

- ✔ We have only used Tweets posted in English.
- ✔ Removing the Null values, Emoji, number and username from the content column of the data.

2

- ✔ Using the inbuilt NLTK library functions we removed the stop words.
- ✔ We used lemmatization to make the model more accurate



Sentiment analysis

1

Finding Sentiments of Tweets

- We have used TextBlob library for sentiment calculation.
- It takes filtered tweet as input and finds its polarity on a scale of −1 to 1.
- Based on polarity sentiment is assigned.

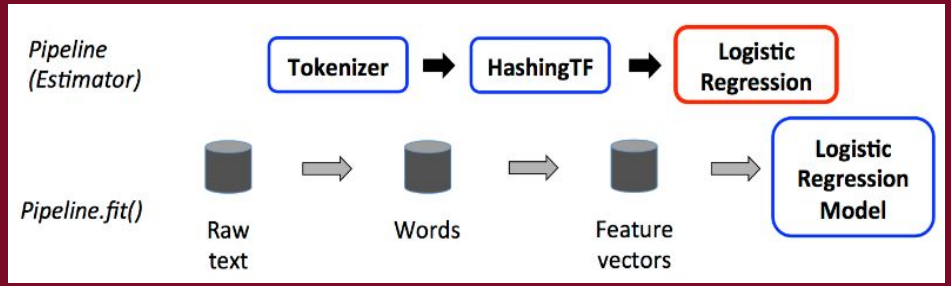
Polarity	Sentiment value	Sentiment
0	0	Neutral
(0,1]	1	Positive
[-1,0)	2	Negative

Sentiment analysis

2

Sentiments Classification

- Logistic Regression and Naïve Bayes classification algorithms are used to classify sentiments.
- Accuracy of Logistic Regression and Naïve Bayes is 89% and 71% respectively.



Evaluation

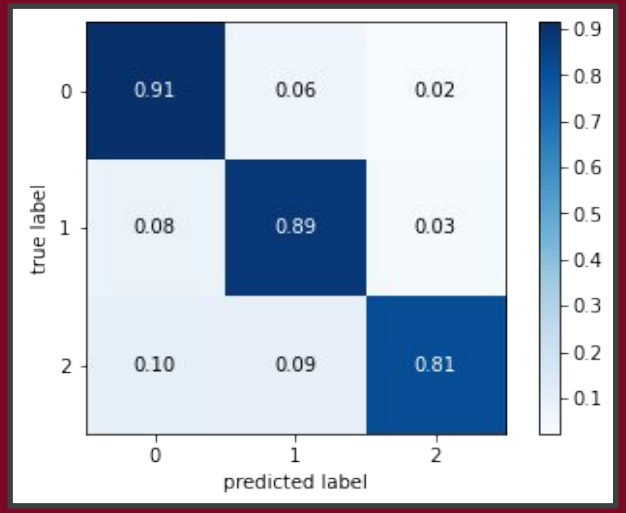
	precision	recall	f1-score	support
0.0	0.89	0.91	0.90	106638
1.0	0.91	0.89	0.90	109564
2.0	0.81	0.81	0.81	31817
accuracy			0.89	248019
macro avg	0.87	0.87	0.87	248019
weighted avg	0.89	0.89	0.89	248019

Logistic Regression

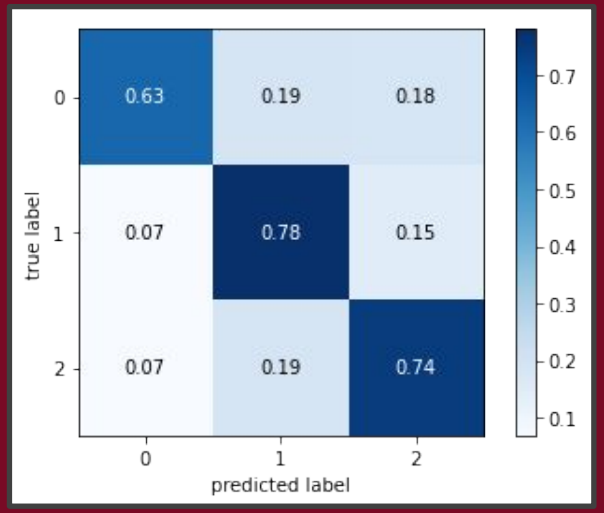
	precision	recall	f1-score	support
0.0	0.87	0.63	0.73	106638
1.0	0.77	0.78	0.77	109564
2.0	0.40	0.74	0.52	31817
accuracy			0.71	248019
macro avg	0.68	0.72	0.68	248019
weighted avg	0.77	0.71	0.72	248019

Naïve Bayes

Evaluation



Logistic Regression



Naïve Bayes

Queries Evaluation

1

Top N likeCount on same dataset

We have used **df.top(n, key)** query to find top N like counts. It returns most liked tweet first.

It require dataframe to be stored in memory. For that reason, we created RDD of necessary columns only.

Top	Query time
1	8.243618249893188
100	8.63142704963684
1000	9.92312479019165

Queries Evaluation

1

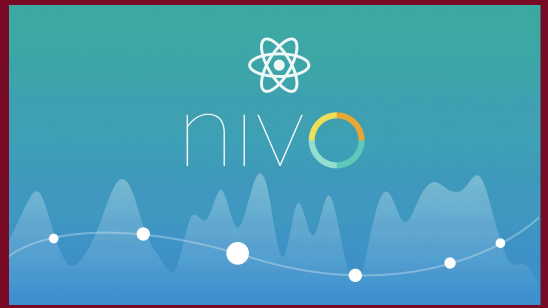
Top 10 likeCount on different dataset

To understand pyspark sql queries capabilities and limitations we ran same query on different data size like 60mb, 350mb, and 3gb.

Data Size	Query time
60mb	3.147566556930542
350mb	8.582376956939697
3.0gb	115.25807619094849

Visualization Dashboard: technology stack

- The Frontend is built using HTML, CSS, JavaScript and React JS.
- Various libraries of React JS like react-tables, nivo are also used for effective visualization.



Visualization Dashboard Demo

FIFA WORLD CUP
Qatar2022

Related Work

- Behavioral analysis on World Cup data
- Data Pre-processing
- Sentiment Analysis
- Data Analysis and Visualization

Conclusion

1. Supervised ML models were used to learn sentiments from the custom dataset and the accuracy of Naïve Bayes and Logistic regression was evaluated and compared

   The interactive dashboard tool will be evaluated based on query response times to check for scalability.

2. In the future, Deep learning model could be implemented to improve the accuracy of sentiment analysis, Transformer models like BERT and ELMO

   Live tweet visualization could be the next feature to add to this dashboard tool which will visualize tweets based on specified timeframes

Interesting fact!

Final: ARG vs POR?

Who do you think will win?

FIFA WORLD CUP
Qatar2022

Thank you

FIFA WORLD CUP
Qatar2022

Data Pre-processing

2

Spark

We used spark to process the data which we will take from the HDFS. And pyspark for further processing the data

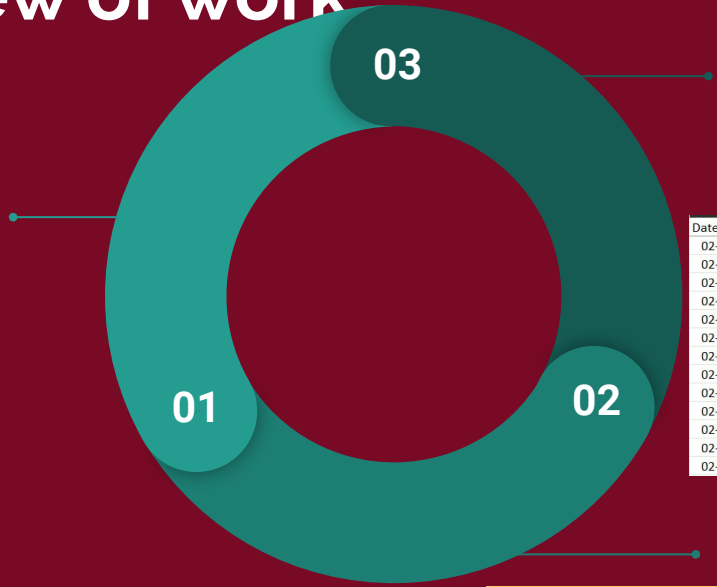
Data Pre-processing

1

Hadoop

We will store the data at the Hadoop cluster. That will replicate the data into multiple datanodes.

Overview of work done



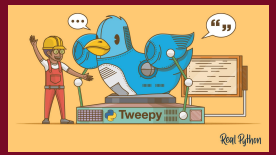
Twitter Features

Twitter is a big source of data containing many features which is useful for Visualizing and Sentiment Analysis

Sample Twitter Dataset

Our dataset will consist of features like date, location, url, text, etc. which are relevant for visualization and sentiment analysis

Date	Source	len	Orig_Twee	Tweet	Likes	RTs	Hashtags	UserMenti	UserMenti	Name	Place
02-07-2018 1.33	Twitter for	95	RT @Manl	Commiser	0	790	ESP,World	Manchest	ManUtd,D	Luzman Na	Gerik, Pera
02-07-2018 1.33	Twitter for	103	RT	Lightsaber	0	4	PowerByE	Johanna Xi	mi_xiuche	Liz JimÃ©nez	Gante
02-07-2018 1.33	Twitter for	135	@FIFAWo	please play	0	1	EXO,Worlc	FIFA Worlc	FIFAWorld	? meaw ??	Bangladesl
02-07-2018 1.33	Twitter for	108	RT	Artificial Ic	0	5	PowerByE	Johanna Xi	mi_xiuche	Liz JimÃ©nez	Gante
02-07-2018 1.33	Twitter for	109	RT @LFC:	Dejan Lovr	0	504	CRO,DEN,\	Liverpool F	LFC	lj0615 ???:	A+ THAILA
02-07-2018 1.33	Twitter for	106	RT @Pumu	on Penaltie	0	1	ESP,RUS,M	Liz Aceved	Pumulo86	Slimboy fa	Dallas, TX
02-07-2018 1.33	Twitter We	100	Germany	Germany C	0	0	WorldCup			Kamal Mus	Bengaluru,
02-07-2018 1.33	Twitter for	138	RT	Kasper Sch	0	2199	Manofthel	FIFA Worlc	FIFAWorld	Ana C. of T	Brasil.
02-07-2018 1.33	Twitter for	138	RT	completed	0	14	Spain,ESPF	Jason Fost	JogaBonit	The US of	United Sta
02-07-2018 1.33	Twitter for	116	RT	quarterfina	0	544	WorldCup,	HNS | CFF	HNS_CFF	EvilX	BKK Thaila
02-07-2018 1.33	Twitter for	140	RT @Toro	So many fa	0	2	WorldCup	TorontoSt	TorontoSt	Michau va	Toronto, C
02-07-2018 1.33	Twitter for	141	One of the	One of the	0	0	WorldCup,Subasic			Fitz 54	Republic o
02-07-2018 1.32	Twitter for	245	@FIFAWo	What we h	0	0	PowerByE	FIFA Worlc	FIFAWorld	StayStrong	Narnia,Lal



Tweepy

Tweepy twitter API to collect tweet data trending hashtags: #FIFAWorldCup, #Qatar2022, #WorldcupQatar2022, etc

References

[1] Most Popular Sport by Country 2022. Retrieved October 25, 2022 from https://worldpopulationreview.com/country-rankings/most-popular-sport-by-country
[2] FIFA Financial Report 2018. Retrieved October 25, 2022 from https://digitalhub.fifa.com/m/337fab75839abc76/original/xzshsoe2ayttyquuxhq0-pdf.pdf
[3] How Twitter is counting down to the World Cup. Retrieved October 25, 2022 from https://marketing.twitter.com/en_gb/insights/how-twitter-is-counting-down-to-the-world-cup
[4] Lucas GM, Gratch J, Malandrakis N, Szablowski E, Fessler E, Nichols J. GOAALLL!: Using sentiment in the world cup to explore theories of emotion. Image and Vision Computing. 2017 Sep 1;65:58-65. DOI: https://doi.org/10.1016/j.imavis.2017.01.006
[5] 4 Benefits of Twitter Sentiment Analysis for Your Business | Scraping . Retrieved October 25, 2022 from https://scrapingrobot.com/blog/twitter-sentiment-analysis/
[6] https://www.kaggle.com/datasets/rgupta09/world-cup-2018-tweets
[7] Discover the 4 V. Retrieved October 25, 2022 from https://opensistemas.com/en/the-four-vs-of-big-data/
[8] Taneja S, Taneja M. Big Data And Twitter. International Journal Of Research In Computer Applications And Robotics. Vol. 2014;2:144-50. ISSN: 2320-7345
[9] Various ways to evaluate a machine learning model. Retrieved October 25, 2022 from https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15
[10] Insights. Retrieved October 25, 2022 from https://marketing.twitter.com/en/insights
[11] White Fifa World Cup 2022 Logo Wallpaper. Retrieved October 25, 2022 from https://wallpapers.com/wallpapers/white-fifa-world-cup-2022-logo-wah9nrnsy9b12kf4.html
[12] Tweepy – Real Python. Retrieved October 25, 2022 from https://realpython.com/twitter-bot-python-tweepy/
[13] Beginners Guide to Data Preprocessing in Machine Learning. Retrieved June 15, 2020 from https://thelastbyteblog.wordpress.com/2020/06/15/beginners-guide-to-data-preprocessing-in-machine-learning/
[14] Sentiment Analysis. Retrieved from https://dribbble.com/shots/4226968-Sentiment-Analysis
[15] World Happiness Dashboard in Tableau. Retrieved Jan 29, 2021 from https://towardsdatascience.com/world-happiness-dashboard-in-tableau-4dc504212288
[16] Icon Logo, Twitter logo, Twitter logo, blue, social Media png | PNGEgg. Retrieved October 25, 2022 from https://www.pngegg.com/en/png-bbtjg
[17] Pipeline image from https://spark.apache.org/docs/latest/ml-pipeline.html

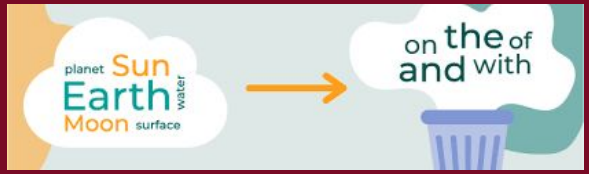
Data Pre-processing

1

Data Cleaning

Tweets of only English language are processed. Removing the Null values, Emoji, number and username from the content column of the data.

Data Pre-processing



2

NLTK

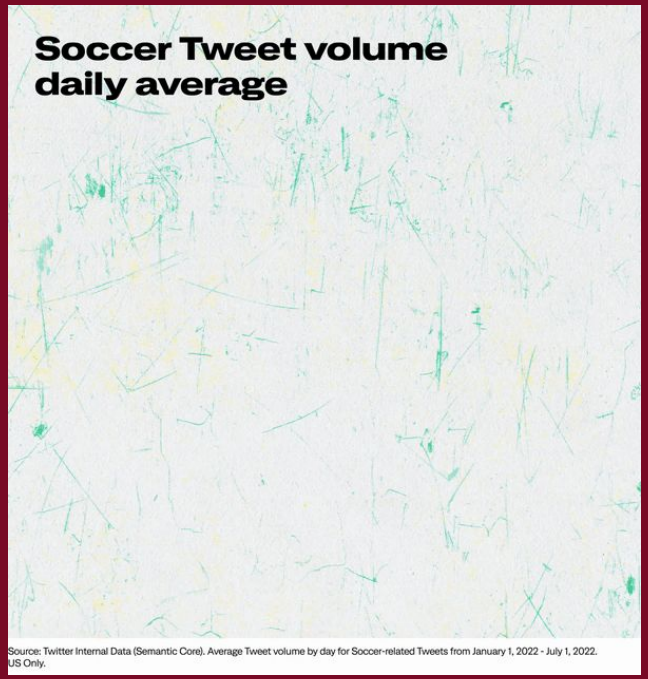
Using the inbuilt NLTK library functions we removed the stop words. And also, the lemmatization to make the model more accurate

Motivation

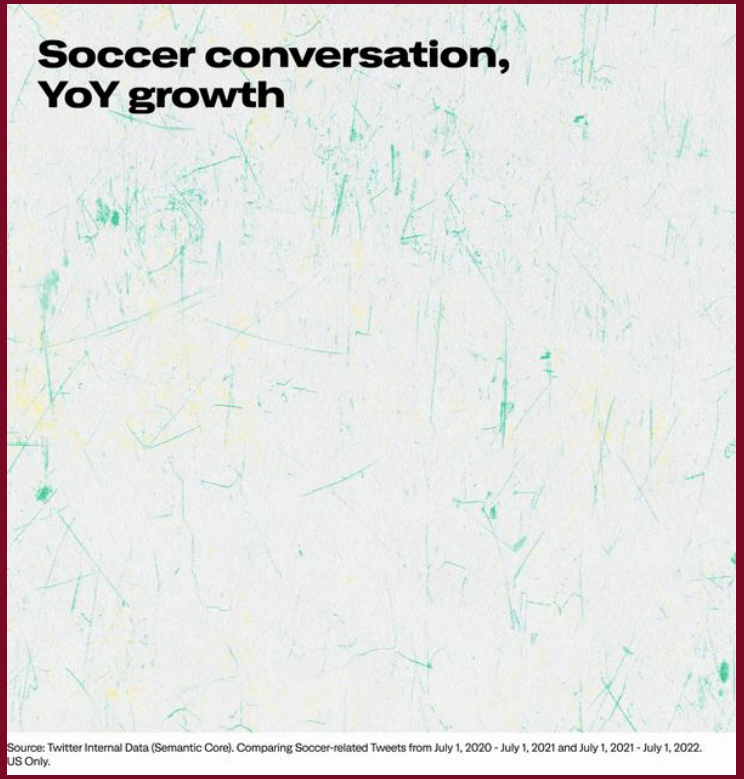
1. Visualizing twitter data can benefit the marketers to analyze reports, interests, evaluate performance of strategies

2. Sentiment analysis helps to find hidden patterns such as brand perception which is a massive factor for large businesses

Twitter Trends

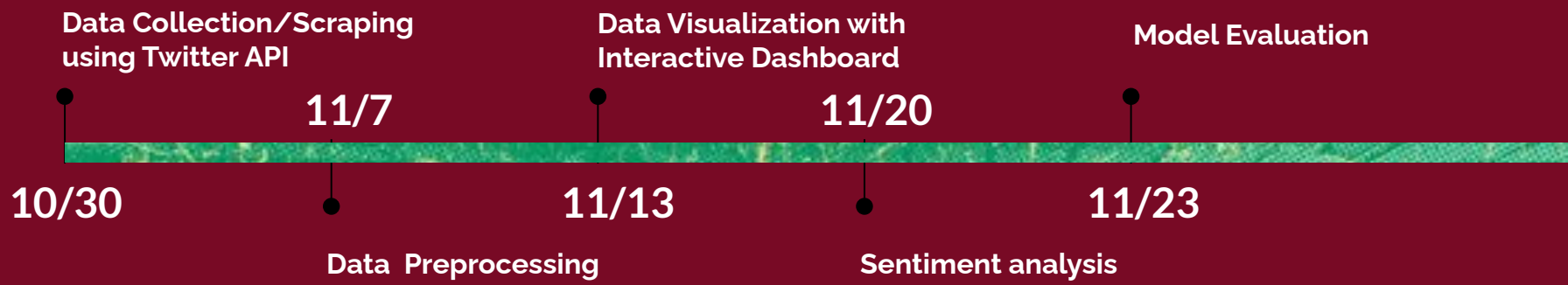


Source: Twitter Internal Data (Semantic Core). Average Tweet volume by day for Soccer-related Tweets from January 1, 2022 - July 1, 2022.
US Only.



Source: Twitter Internal Data (Semantic Core). Comparing Soccer-related Tweets from July 1, 2020 - July 1, 2021 and July 1, 2021 - July 1, 2022.
US Only.

Vision


Data Collection/Scraping using Twitter API
11/7
Data Visualization with Interactive Dashboard
11/20
Model Evaluation
10/30
11/13
11/23
Data Preprocessing
Sentiment analysis

Meet Our Team