Twitter Data Analysis

CS 226: BIG-DATA MANAGEMENT, Fall 2022

Table of Contents

FIFA WORLD CUP
Qatar2022

Background

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

2 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

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

WHY Twitter?

➔ 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

WHY Twitter?

➔ 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

Increase in tweets

425%

World cup tweet trend from February to March 2022

WHY Twitter?

➔ 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

41M

Tweets in UK since the beginning of 2022

Increase in tweets

425%

World cup tweet trend from February to March 2022

WHY Twitter?

➔ 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

41M

Tweets in UK since the beginning of 2022

Increase in tweets

425%

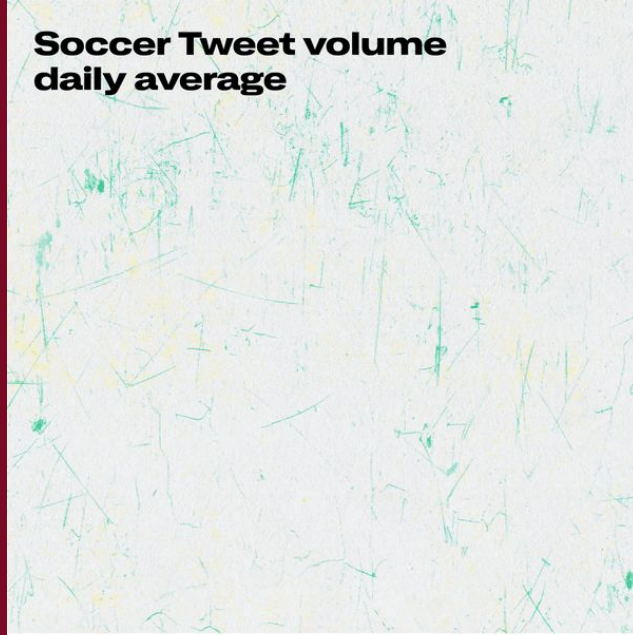**World cup tweet trend from February to March 2022**
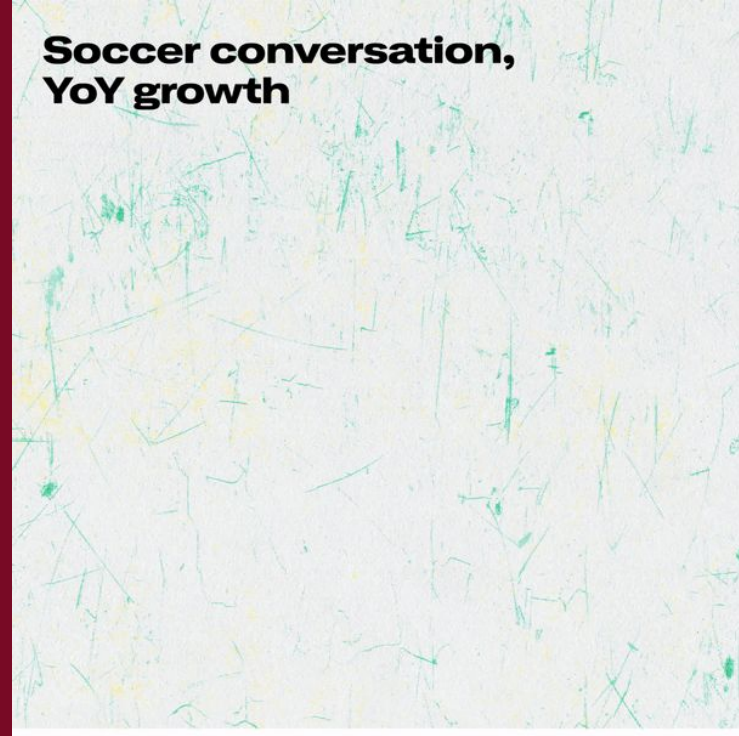
People on twitter

53%

Fans of the World Cup

Twitter Trends

Soccer Tweet volume daily average

Soccer conversation, YoY growth

Dataset

03

01

02

Twitter Features
Twitter is a big source of data containing many features which is useful for Visualizing and Sentiment Analysis
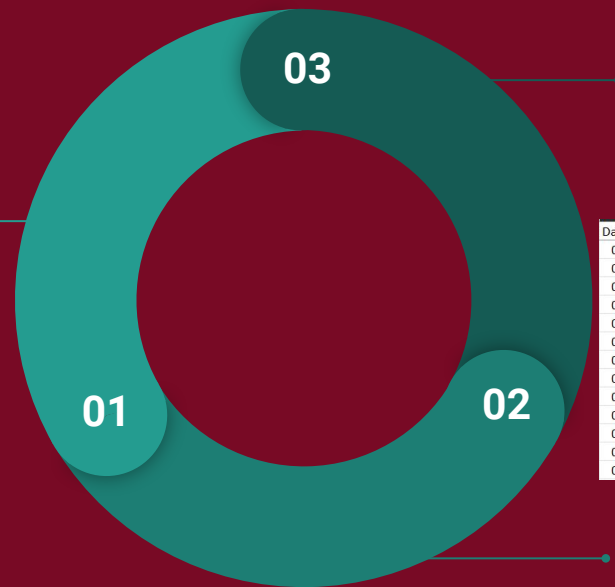
Sample Twitter Dataset
Our dataset will consist of features like date, location, url, text, etc. which are relevant for visualization and sentiment analysis

Date	Source	len	Orig_Twee	Tweet	Likes	RTs	Hashtags	UserMenti	UserMenti	Name	Place
02-07-2018 1.33	Twitter for	95	RT @Man	Commiser	0	790	ESP,World	Mancheste	ManUtd,D	Luzman Na	Gerik, Pera
02-07-2018 1.33	Twitter for	103	RT	Lightsaber	0	4	PowerByE	Johanna X	mi_xiuche	Liz JimÃ©nez Gante
02-07-2018 1.33	@FIFAWo	135	please play	0	1	EXO,Worlc	FIFA Worl	FIFAWorld	? meaw ??	Banglades
02-07-2018 1.33	Twitter for	108	RT	Artificial Ic	0	5	PowerByE	Johanna X	mi_xiuche	Liz JimÃ©nez Gante
02-07-2018 1.33	Twitter for	109	RT @LFC:	Dejan Lovr	0	504	CRO,DEN,\	Liverpool I	LFC	lj0615 ????	A+ THAILA
02-07-2018 1.33	Twitter for	106	RT @Pum	on Penalti	0	1	ESP,RUS,\	Liz Aceved	Pumulo86	Slimboy fa	Dallas, TX
02-07-2018 1.33	Twitter W	100	Germany	Germany (	0	0	WorldCup			Kamal Mus	Bengaluru,
02-07-2018 1.33	Twitter for	138	RT	Kasper Sch	0	2199	Manofthel	FIFA Worl	FIFAWorld	Ana C. of 1	Brasil.
02-07-2018 1.33	Twitter for	138	RT	completed	0	14	Spain,ESPI	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 Worl	FIFAWorld	StayStrong	Narnia,Lak

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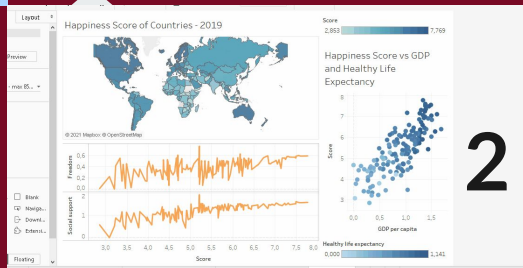
Process

1

Data preprocessing

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

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.

2

3

Andy Williams
CTO, Market Tech

Does anyone know when the shuttles leave this morning?

0   3

Sentiment Analysis

We will be using Machine learning model on Tweets to classify the Sentiments such as SVMs, Random Forest, Naive Bayes, etc.

Outcome

1 Visualization Dashboard

We plan to display top N tweets with respect to clicks, likes, location, etc. We will also try to plot a point map of tweets with respect to locations around the world

2 Sentiment Analysis

We plan to use NLP techniques to extract polarity of sentiments using text present in tweets. We will assign sentiment score to tweets using custom libraries and train our model
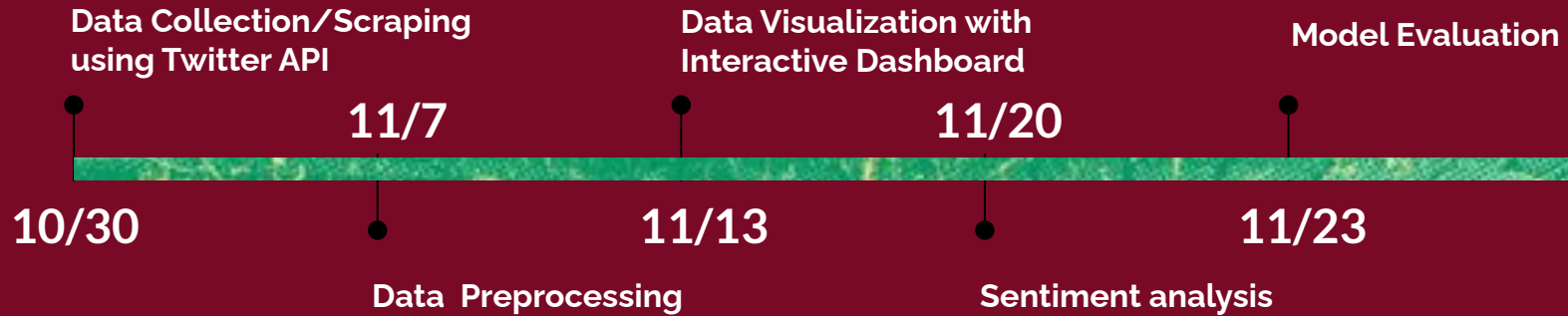
Evaluation

Visualization Dashboard

1

❖ Query Response Time
❖ Compare and contrast of query response time for various queries and data sizes

Sentiment Analysis

2

❖ Plot Confusion matrix and the ROC (Receiver Operating Characteristics) curve
❖ Accuracy, Precision, Recall, and F1

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

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