Twitter Data Analysis: FIFA World Cup Qatar 2022

Group Name: Fantastic4

Vineet Dhaimodker
862255153
vnaiq001@ucr.edu

Vishv Patel
862322103
vpate062@ucr.edu

Mihir Patel
862324469
mpate125@ucr.edu

Yash Gandhi
862324764
ygand001@ucr.edu

ABSTRACT

FIFA World Cup is the most viewed event with about 3.5 billion followers worldwide. The 2018 FIFA World Cup generated more than 5 billon USD of revenue in total. So far there have been a total of 41 million tweets related to soccer in UK alone, since the beginning of the year. There has been an 425% increase in world cup tweets from February to March. This goal of this project is to build an interactive dashboard tool to visualize the twitter data about the Football Worldcup 2022 held in Qatar and thereby enable businesses to reveal crucial insights. This project uses a custom dataset collected using relevant hashtags over the period of January 2022 to November 2022 having almost 5 million tweets. The data is cleaned and processed using NLTK and Textblob, then used to perform sentiment analysis. The project employs supervised Machine learning techniques such as Naive Bayes and Logistic Regression to predict sentiments of a given tweet. We found that Logistic Regression outperformed Naïve Bayes with an accuracy of **89%**.

1. Introduction

1.1 Motivation

Sporting events like the football World Cup bring forth strong emotions of viewers and are a reliable source to explore how sentiments affect businesses. Visualizing twitter data can benefit the marketers to analyze reports, interests, evaluate performance of strategies. Visualizing FIFA world cup tweet data will help businesses improve marketing strategies and better reach their target demographic. On the other hand, Sentiment analysis helps to find hidden patterns such as brand perception which is a massive factor for large businesses.

1.2 Problem Statement & Challenges

Due to the influence of twitter on the lives of millions of people, businesses use it for marketing, growth and development and take feedback from their followers for consistent improvement. The aim of this project is to project is to visualize interesting statistics related to FIFA World Cup 2022 in hope to reveal insights. Further, we analyze sentiments of fans realize brand perception and emotion for the game. The result of this project could also be used to track statistics of target demographic for brands in industries like sports apparel, clothing, travel, supplements, etc. to improve product quality and customer satisfaction.

The challenges for this project were as follows:

- Gathering relevant data: The target data for this project is very specific to FIFA Worldcup 2022 therefore it had to be collected manually from scratch. The sheer volume of target data became a hurdle as the twitter API has a limit of 900 requests/15-minutes. Also, volume of twitter users who are soccer fans makes it tough to get data relevant to a particular topic, which is where relevant hashtags come into picture.
- Tweet Veracity: Twitter supports more than 100 languages which is a challenge as we perform sentiment analysis for English tweets. The raw tweets include emojis, hashtags, user mentions, special characters, etc. These may hinder results while performing sentiment analysis. Dropping all tweets in other languages, removing unnecessary text for such a large dataset poses a challenge.
- Custom function in spark session: Sentiment assignment to the tweets was challenging considering understanding on how to use a user defined function (udf) and registering it with the spark session for future use.
- Big data query in Dashboard: To make the dashboard interactive for user, we need to display output depending on user requirement and since we have a large volume of data, normal querying techniques might not scale well. To overcome this we will have to use pyspark for the query processing in the back end.

Following sections include the survey of past work and relation to our project, data gathering, preprocessing, cleaning of the data, ml models, model evaluation, dashboard, and query evalution.

2. Literature Survey

2.1 Behavioral analysis on World Cup data

In this section, we will look at previous studies which of behavioral analysis on World cup data. The paper [2] is a study which uses sentiment analysis on 2014 World cup tweets showing insights that make matches interesting. The paper [3] is a big data analysis study of change in emotions after goals being scored of US sports fans during the World Cup 2014. The paper [4] is a Big Data Analysis of English National Team Fans During the 2018 FIFA World Cup showing concepts of Basking in Reflected Glory and Cutting Off Reflected Failure during the match. We take inspiration from these papers to use the 2022 Football world cup tweets data as our target dataset to perform relevant analysis.

We take a look at a paper that uses sentiment analysis on tweets posted during the 2014 World cup to explore theories regarding the emotions of fans[2]. The authors argue how sports have served as a source of data to understand the emotions and cognitive functioning of the human brain and why they focus on the emotions of fans of World Cup soccer by providing relevant statistics. This paper uses the 2014 world cup twitter corpus captured using Twitter Streaming API and the Twitter4J Java library. The tweets recorded were started 15 minutes before each match began till the end of the game. For sentiment analysis, the authors used a variation of the SemEval Twitter sentiment analysis challenge submission which was based on lexicon-derived features to show the distribution of emotions using several statistics. The novelty of this paper lies in the new approach presented by the authors for the uncertainty of the outcome hypothesis. Unlike ticket sales or TV viewing, emotional content in tweets was effectively used to show the extent to which a match becomes uninteresting and the tweets during the match, were related to a lower percentage of positive tweets. Thus, the results show how sentiment analysis could give insights into what makes sports matches interesting. In future work, the authors mention that the geographic information of Twitter users could be used to determine which team the tweet was meant to support. We plan to incorporate similar ideas into our project.

Another paper pertaining to the big data analysis of US sports fans during the World Cup 2014 shows the change in emotions after goals being scored [3]. This paper shows how fear and anger were the commonly displayed emotions during matches involving the home country and how these emotions grew when the opponent team scored and reduced when the home team scored. The authors have done an excellent job of explaining how fans show emotions during games during the world cup using the example of the famous 7-1 loss faced by Brazil against Germany in the 2014 World Cup semi-final game. The paper examines how U.S. sports fans reacted to matches in which the U.S. National Soccer team competed during the FIFA World Cup 2014. The authors have done a great job in making the data cleaning process clear to the readers, wherein they have used steps like noise cleaning, tokenization, stop word & emoticon removal, and stemming & lemmatization. Each tweet is broken down to portray emotions like Anticipation, Fear, Anger, Surprise, Sadness Joy, and Disguise, using lexicons. The authors tabulate the results of the number of words with specific emotions and total tweets for games played by the US team against Portugal, Germany, Belgium, Nigeria, and Colombia. The authors present three major findings. First, negative emotions increased while the home team was losing or drawing and decreased when they were winning. Second, the authors were able to show that, their approach provides live interactions from fans, and can be an alternative to controlled experiments i.e natural experiments could be used to validate results. Third, big data analysis helped authors to analyze large number of tweets and further examine the natural emotion of fans in real life. The limitations of the approach used in this paper are twofold, one is that they did not consider the semantic meaning of the whole tweet and the second is that the sentiment lexicon fails to cover domain knowledge and extract word meaning in a specific scenario.

The Big Data Analysis of English National Team Fans During the 2018 FIFA World Cup [4], shows how English fans performed Basking in Reflected Glory (BIRG) and Cutting Off Reflected Failure (CORF) during their matches against Croatia and Columbia. The authors collected the data using the streaming API using keywords like "FIFA", "WORLD CUP" and location steps from "England", and "United Kingdom". Tweets were collected in a 3-hour window from the beginning of the match till 30 minutes after the end of the match. The authors have used a comprehensive method for identifying which team the user is supporting using keywords and phrases and represented it as 0 for no national identification, 1 for national identification with England, and 2 for national identification with Croatia or Colombia. The authors have also devised a machine learning algorithm to not only decipher differences in tweet valence at different times during the game but also to analyze emojis and also analyze those depictions. The authors use Term Frequency-Inverse Document Frequency (TFIDF) to improve query retrieval. The authors use mainly 3 models for predictions: logistic regression, Light Gradient Boosting, and XGBoost along with 10-fold cross-validation and present results that show that these models have more than 90% accuracy in team identification, and 89% accuracy in Nation identification.

2.2 Data Preprocessing

In this section, we observe previous studies which showcase data preprocessing techniques used in similar applications. In the paper [5], authors have cleaned the data by removing URL, hashtags, and replacing special symbols to classify sentiment of tweets. In paper [6], authors handle big data using Hadoop and used NLTK for tokenization, POS tagging, and lemmatization to process the data to find sentiments. Paper [7], used NLTK library to clean data by removing linguistic mistakes such as repeated letter/word, noise, POS tagging, and stop word removal. The paper [8], gives the insight of lexical-based approach to label the sentiment of each text. The paper [1] discusses common approaches to manage big data such as Distributed Parallel Processing (Apache Hadoop and MapReduce), In-Memory Databases, NO SQL Databases, etc. along with how they could be used for Twitter data and compares data analytics tools like Twitter Counter, Twitonomy, Twtrland, etc. which help to analyze clients' experience. These tools give us an idea about types of data we need to analyze. We implemented few of these techniques for cleaning our custom dataset of football tweets by removing meaningless words, hashtags, user mentions, and emojis. We also used NLTK libraries for removing stop words, lemmatization, and tokenization. Finally, we labeled sentiments to tweets using Textblob library.

In the paper [5], a method for sentiment analysis of tweets is proposed. This approach consists of three main steps: subjective classification, semantic association, and polarity classification. Extracted tweets from Twitter and manually labeled them into three groups: positive, negative, and neutral. After extraction, the tweets were pre-processed and converted into a formal machine-understanding language. The sentiment classifier predicted tweets as positive, negative, or neutral by classifying emotions from the sentiment dictionary. The analytical model is performed on the pre-processed data; the paper has shown insights into steps to pre-process the data. The purpose of preprocessing is to process the raw tweets and display them in a clean form to improve the machine's understandability of the text. This includes removing URLs and hashtags, replacing special symbols, removing repeated characters, expanding abbreviations and acronyms, and capitalizing topics.

In the paper [6], they processed a huge amount of tweets for sentiment analysis. As the number of users on social media are increasing rapidly, the data is also growing and they have considered tweets for sentiment analysis. To handle the enormous volume and variety of data, Hadoop/MapReduce is taken into consideration. For pre-processing the data, they have applied many methods to transform raw data into clean data, such as removing stop words, URLs, numbers, and hashtags. Further, to have a basic text which can be easily understood by the machine, they applied NLTK processing libraries which include tokenization, POS tagging, and lemmatization. To classify sentiment as positive, negative, and neutral, they used the Naive Bayes model.

The paper [7], presents the sentiment analysis on Twitter data. They collected the tweets using the Twitter API with the keywords. The tweets are in raw form, so NLTK libraries are applied to remove linguistic mistakes such as repeated letter/word, noise, POS tagging, and stopword removal. Trained multiple models with different subsets of training data, namely Naive Bayes, SVM, and Ensemble. The Ensemble model outperforms all other models in that it includes the extremely randomized trees classification.

The paper [8] describes a lexical-based approach for extracting sentiment from text. The idea is a semantic orientation calculator using a word thesaurus with semantics. Semantic orientation can be either positive, negative, or neutral. To compute arbitrary text alignment, adjective-adverb combinations are extracted with sentiment alignment values. These are converted into a single score by aggregating all values.

The paper [1] discusses Twitter as a perfect example of social media as big data. The author has done so by providing statistics such as Twitter had 250+ million active monthly users sending 500+ million tweets each day with each tweet having a maximum of 140 characters and supporting 35+ languages. This showed that Twitter could be used to determine the root cause of global events and thereby be used by businesses and celebrities to promote themselves to reach their followers. The author made it clear that the volume, velocity, veracity, and variety of Twitter data fit the description of big data. The author has provided common approaches to manage big data such as Distributed Parallel Processing (Apache Hadoop and MapReduce), In-Memory Databases, NO SQL Databases, etc. along with how they could be used for Twitter data. By analyzing tweets, finding keywords, and using map-reduce we can filter out trending Twitter topics. Similarly, computing a score for keywords about brands will help find the brand perception i.e. Sentiment Analysis. The paper describes twitter data analytics tools like Twitter Counter, Twitonomy, Twtrland, etc. which help to analyze clients' experience, recommend the perfect time to tweet, and suggest people to follow. The paper also provides a - comparison of Twitter Analytics Tools with respect to features and whether they are free or paid.

2.3 Sentiment Analysis

In this section, we review Sentiment analysis techniques that have been used in similar applications. Paper [9] gives insights into feature selection and semantic classification techniques. Furthermore, supervised learning classifiers like Naïve Bayes, Support Vector Machine (SVM), and Decision tree are compared. Paper [10] focuses on feature extraction techniques like Bag of Words, and n-gram. Paper [11] has explained the difference between unsupervised and supervised learning for Sentiment Analysis and WordNet is used to find the polarity of the texts for sentiment analysis. Paper [12] compares supervised learning, unsupervised learning, and hybrid learning when finding semantics and uses Naive Bayes, Maximum Entropy, and SVM models. Taking inspiration from these papers we have used Naïve Bayes and Logistic Regression to classify sentiments. Also, we have planned to use N-gram for making input more meaningful.

Paper [9] gives an in-depth knowledge of Sentiment Analysis (SA) techniques along with categorizing them based on fields namely Emotion Detection (ED), Building Resources (BR), and Transfer Learning (TL). Feature selection techniques like Point-wise Mutual Information (PMI), Chi-square, and Latent Semantic Indexing are explained in detail. Semantic Classification Techniques like the machine learning approach, lexicon-based approach, and hybrid approach are discussed in detail. Supervised Learning classifiers like Naïve Bayes Classifier, Maximum Entropy Classifier (ME), Support Vector Machines Classifiers (SVM), Decision tree classifiers, and others are explained and compared.

Paper [10] discuss the Opinion Mining and Sentiment Analysis (OMSA) challenges with big data. Explains the rising of data and why Twitter is a great source for big data. Apart from Twitter, data from Amazon, Youtube, and Tripadvisor are taken to show the diverse applications of OMSA. For the semantic analysis 3 approach are explained naming Keyword-based Classification, Lexicon Based Classification, and Machine Learning- Based Approach. Features extraction techniques for the different data source is discussed like Bag of Words, n-gram, Bag of Concepts, based on lexical features, etc. To show the effectiveness and real applications many machine learning algorithms such as Random Forest, SVM, K-nearest neighbor, etc are used. Financial and sports sectors and how the OMSA technique can be useful in them are shown.

Paper [11] has described the challenges and needs for performing sentiment analysis. Explained the difference between unsupervised and supervised learning and details of their latter selection. Discuss approaches for extracting text and why they chose the Unigram model. Supervised techniques like Naive Bayes, Maximum entropy, and Support vector machine have been used in this paper for training and classification of the texts into semantics. Flow and stepwise implementation of the whole process with pseudo-code is shown. The evaluation is done using Precision and Recall. After training and classification, semantic analysis derived from WordNet is used to find synonyms and the polarity of the texts.

Paper [12] focuses on the comparison of supervised learning, unsupervised learning, and hybrid learning when finding semantics. Uses Naive Bayes, Maximum Entropy, and SVM models. Along with this, it also finds sentiments on Document-level as well as Sentence-level sentiments. Features extracted using Unigrams, Bigrams. Foundings combining various features improved the accuracy of machine learning algorithms.

2.4 Data Analysis and Visualization

In this section, we review Data analysis and visualization applications on similar data. The paper [13] emphasize the importance of analyzing the Twitter location with sentiment analysis on the 2016 US elections data by creating web page to visualize this data using plots and graphs, and maps. The paper [14] presents a web tool TweetViz; to visualize specific user behavior change over time using user hashtag distribution and temporal distribution of keywords. The paper [15] present a framework for real-time analysis of Twitter data to detect relevant topics by selecting the K most relevant terms with highest likelihood of appearance. We plan to incorporate the feature of selecting top k tweets with respect to fields such as like count, retweet count, etc. and plots graphs with top k countries with highest tweets. We also plan to use location to plot an expandable map.

A lot of user data is generated by Twitter as the number of users has tremendously increased. The authors in this paper [13] are trying to emphasize the importance of analyzing the Twitter location data with sentiment and behavior analysis. To demonstrate the importance, the authors have chosen to perform data analysis on the 2016 US elections. The authors can perform a fine-grained analysis of the tweets using the location data, which enables them to get more accurate results of the user’s sentiment and behavior analysis. A web page is created to visualize the data using plots and graphs, but more importantly, they use maps that use the location information from the metadata of the tweets which gives a more sophisticated and detailed analysis of the subjectivity and polarity during the elections.

This paper [14] presents a web tool to visualize Twitter data. The TweetViz visualizations are of two types. One of them is user-centric and its main goal is to analyze specific user behavior. This is achieved by plotting charts that show the number of tweets the user posts on a daily basis and analyzing the change in user activity. The user hashtag distribution is visualized, and the temporal distribution of keywords or hashtags is visualized all giving interesting insights about the user. The authors also propose an approach of visualizing topic distribution in a set of tweets over some time interval. Latent Dirichlet Allocation algorithm is used to achieve it. Using visualization representation of topic distribution provides information about how user interests change over time.

The authors of this paper [15] present a framework for real-time analysis of Twitter data which is helpful to detect relevant topics discussed by the users. They propose improvements to the existing method which is Soft Frequent Pattern Mining (SFPM). Given a set of tweets, the initial job is to select the K most relevant terms, which are terms with the highest likelihood of appearance in the current set of tweets. This method works for static data, but when it comes to live detection, the Twitter live Detection algorithm uses the likelihood ratio with a measure to consider the importance of relevant terms in the current set of tweets.

2.5 Categorization

The below Table shows the categorization of the papers we studied as a part of our Literature Survey.

Reference Number	Category	
2, 3, 4	Previous work on Twitter and World Cup data	
1, 5, 6, 7, 8	Data Preprocessing	
2, 3, 4, 9, 10, 11, 12	Sentiment Analysis	
13, 14, 15	Data Analysis and Visualization	

3. Analytical Framework

3.1 Overview

This section discusses about the steps we used to collect the raw data, transform the data based on the application requirements and finally perform relevant experiments. We started by using snscrape to collect large-scale custom data followed by PySpark for processing the collected data. Tweets were cleaned by performing various operations, extracting meaningful text. Further, we used Textblob library for assigning sentiment based on polarity to prepare training data. We then pipelined TF-IDF transformation with ML algorithms to make analysis scalable. For better visualization, we have been working on various components of the visualization dashboard tool and plan to integrate all components.

Hadoop - Large datasets of gigabytes in size can be processed and stored effectively using Apache Hadoop architecture. Hadoop enables clustering multiple machines to examine big datasets in parallel more than using a single powerful machine for data storage and processing. Thus, we implemented Hadoop to our local machine and stored the data to HDFS (Hadoop Distributed File System). HDFS is distributed file system that functions on common hardware. As we can compare to conventional file systems, HDFS offers higher data speed, superior fault tolerance, and native support for large datasets. And YARN (Yet Another Resource Negotiator) which manages and monitors the Hadoop cluster and resource usage, also it will schedule jobs and task for any callable functions accordingly. [16]

Spark - Big data workloads are processed using Apache Spark, a distributed processing engine. It uses efficient query execution and in-memory caching for quick analytic queries for large size of data. It allows reusing of code across a variety of workloads, including machine learning, batch processing, interactive queries, and graph processing. [17] We will process our data using Py spark which is one of spark model that can be run on python. Pyspark will help to process the large amount of data easily and faster. Further, MLlib of Py spark is deployed to use the machine learning algorithms.

3.2 Data Gathering

Twitter has imposed limitations on the number of requests that can be made using the twitter API, which is why we have used snscrape scraper for social networking services. We extracted tweets form the beginning of 2022 till the end of the first match of the World cup which was held on November 20. We used top hashtags from mainly three categories namely: trending (eg. fifaworldcup, FIFAWorldCup2022, FIFAWorldCupQatar2022, Qatarworldcup2022), teams (eg. threelions, usmnt, blackstars), and matches (eg. qatarvsecuador, qtrecu, porarg). We extracted 13.6 GB of twitter data with almost 5 million tweets.

The total time required to scrape this data was around 43 hours and 18 minutes. The tweet scrapping script generates a new json file of the raw tweets for each month from January to November, for every hashtag. A total of 583 files were generated, i.e., 11 files for each of the 53 hashtags that were used to capture tweets relevant to the world cup

We extracted the data from the twitter with 21 features namely tweetid, content, likes and many more. Afterwards we stored that data into Hadoop cluster where it will make many chunks based on the replication factor. So, our data does not lose while pre-processing on it. Finally, the data was loaded to the spark to perform the pre-processing using the Py spark library.

3.3 Data Preprocessing

The SNScrape returns 21 fields of data, of these we use 15 fields which are useful for visualizations and sentiment analysis. We used fields like date, content, location, retweet Count, like Count, lang etc. We combined multiple json files generated by the SNS scraper to get all tweets from the English language and imported the combined data using Py spark to perform required preprocessing and cleaning.

3.3.1 Data Cleaning

Raw data was cleaned and processed using Spark operations to prepare it for training and visualization. Remove tweets from languages other than English. Remove unnecessary features, keeping only the associated features required for sentiment analysis and for queries like monthly tweets, top N Likes, top N Replies, and location were maintained out of the 15 features that were extracted. Removing duplicate tweets, as tweets were scraped using 53 hashtags, duplicate tweets which consist of common tweetid were removed to make the dataset unique. Tweet ID, date, content, user location, likes, and replies are some of these aspects. Remove records which have null or empty values for columns such as tweet content. Now we removed user mentions, hashtags, emojis, numbers and special characters using regex.

In order to use Natural Language Processing models, raw tweets had to be transformed. The NLTK library, a natural language toolkit that offers libraries and functionalities for NLP, was implemented. Emojis, hashtags, user mentions, numbers, URLs, and punctuation were among the contents that were eliminated. Tokenization was done in order to translate tweet messages into words. Stop words were extracted from the tokenized tweets using the NLTK function. Stop words have been removed from the tweets as they do not add any meaning to sentiment analysis of the tweet content. Lemmatization removes grammatical forms and transforms each word into its original form. Lemmatization was implemented due to its several advantages for the sentiment analysis. Finally, the data is ready for Polarity and Sentiment Analysis using ML models.

3.3.2 Data Transformation

TextBlob library is useful to perform textual data processing. We used it to find polarity of the tweets. We made a user-defined function using PySpark to find sentiment. It takes each tweet as input. TextBlob converts input tweet by performing built-in tokenization and calculates polarity based on sentiment score of each token. Based on polarity, we assign sentiment value to tweets as follows:

Polarity	Sentiment label
neutral	0
positive	1
negative	2

3.4 Data Processing

To make the machine learning models scalable we have implemented Logistic Regression and Naïve Bayes using PySpark MLlib. Also, we are creating pipeline of different stages containing Transformers and Estimators. It includes stages of features transformer like Tokenizer, features extractors like Hashing TF, IDF, and ML algorithms. All transformations and extractors are implemented using PySpark MLlib. TF-IDF will be the input feature to train the ML models. A multiclass classification evaluator is used to find accuracy for the predicted sentiment on test data. Used scikit-learn classification and confusion matrix to generate detailed analysis on model performance. Plotted the confusing matrix using matplotlib.

4. Model Evaluation

The train-test split for all our experiments is 70-30. We have 1 million tweets for training and testing our models. Logistic Regression gave 81% accuracy while Naïve Bayes gave 71% accuracy. This accuracy was achieved by tuning hyper parameter of models [21]. To improve Logistic Regression, we have used regParam = 0.01. This will improve L2 regularization by adding some penalty to loss function. As Naïve Bayes is probabilistic model, and sometimes lead to zero probability. To handle that issue smoothing value of 1 was added. Precision, Recall and F1-score were used to evaluate and understand the performance of models. It was interesting observation that Precision of negative values calculated using Naïve Bayes was only 0.4. Meaning negative sentiments were classified less correctly. Example, model predicted 40 negative values when there were 100 actual negative values. Also, we plotted confusion matrix. It showed that neutral values were classified accurately in Logistic Regression. 91% of neutral tweets were correctly predicted. However only 63% of correct natural tweets were identified in Naïve Bayes. Evaluation results are shown below.

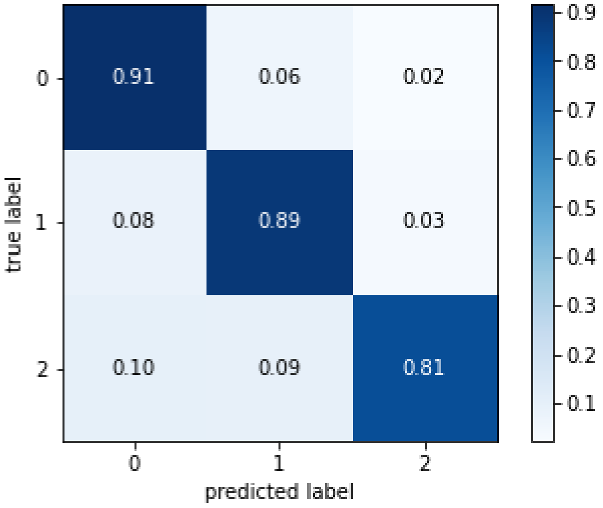


Figure 1: Confusion Matrix of Logistic Regression

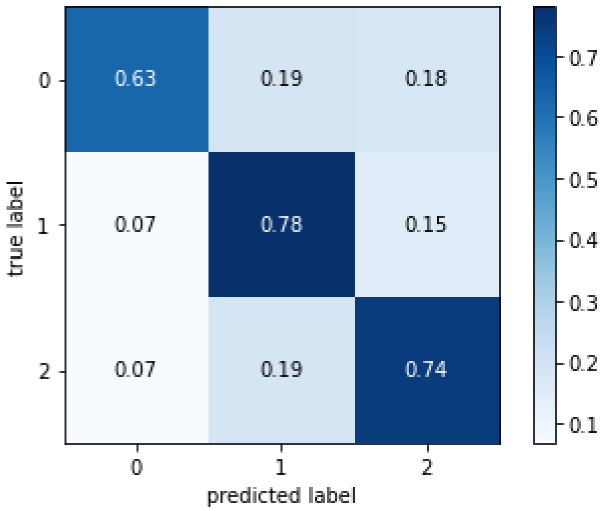
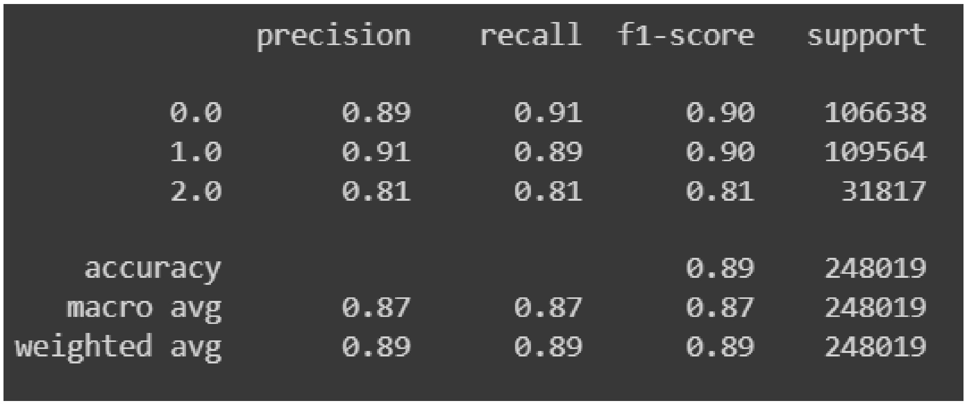


Figure 2: Confusion Matrix of Naïve Bayes

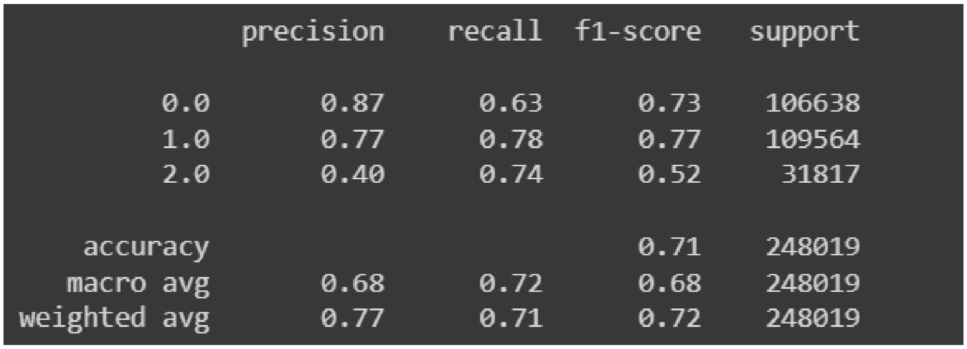


```
              precision    recall  f1-score   support

         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
```

Figure 3: Accuracy, Precision, Recall, and F1-score of 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
```

Figure 4: Accuracy, Precision, Recall, and F1-score of Naïve Bayes

5. Visualization Dashboard tool

The Visualization Dashboard tool displays the analysis in simple easy to understand way. It will use the data to create charts and maps to summarize the analysis and would also be interactive so that user can obtain specific analytics.

5.1. Back-End

We are planning to build custom API for visualization tool. The node and express.js would be used to fetch the data for the frontend based on parameters passed. Here parameters would be selected by users such as top N, sentiments, location, etc. We are planning to use the date parameter for the tweets to fetch number of tweets in certain time frames based on start and end date of time frame provided by user. We will try to display tweet counts and sentiments based on the country location.

5.2. Front-End

The User Interface (UI) consists of a grid view of various charts fit together in a single page to provide an overall summary to the data analysis. The UI also consists of a sidebar which would help the user navigate to each chart / map. This view could consist of an expanded view of the chart/ map. The user can interact with the charts/ map. By hovering over a certain point on the chart or by hovering on a certain country on the world map, a tooltip is provided which contains the analysis summary for that point/ country.

5.3. Visualization

After the data is processed, it is converted into a JSON which is to be passed to the dashboard tool. The dashboard is created using HTML, CSS, Vanilla JavaScript and React js. It consists of multiple reusable components. The main file which is the index.js redirects to the app.js file which contains routes for the landing page, the bar chart containing information about the number of number of tweets and their polarities across the whole year. The second chart contains information about the run times required to query the top 1, 10, 100, 1000 and 10000 tweets about the FIFA World Cup 2022. And third, the world map consists of the number of tweets from different countries.

Details of each of the Components of our Visualization dashboard are as follows:
1. Sidebar.js: This component is used to help the user navigate to different pages like the landing page, and the pages for various charts. This component uses the react-pro-sidebar library, various items from the React Material UI [19] and react-router-dom to link multiple pages.
2. Landing Page: This Page consists of the title and logo
3. Bar Chart (Number of Tweets per month and their polarities): This component uses the Bar Chart component from nivo [18]. The data provided to the chart is analyzed for sentiment, and the chart displays the total number of tweets per month with red, blue and greens as the polarities negative, neutral and positive respectively.
4. Bar Chart (Runtimes for various queries): This component also uses the bar chart component from nivo[18]. The data here consists of the runtime for every query in seconds to calculate the response. The queries are retrieving the top 1 tweet, the top 10 tweets, the top 100 tweets and the top 1000 tweets.
5. Table (Top 10 Most Liked Tweets): This component contains the top ten most liked tweets, the number of likes, its content and a URL to the tweet. It uses react-tables [20]
6. Geography Chart: This component uses the geo map from nivo. It is an interactive map which displays the number of tweets per country using a color scheme to denote a range.

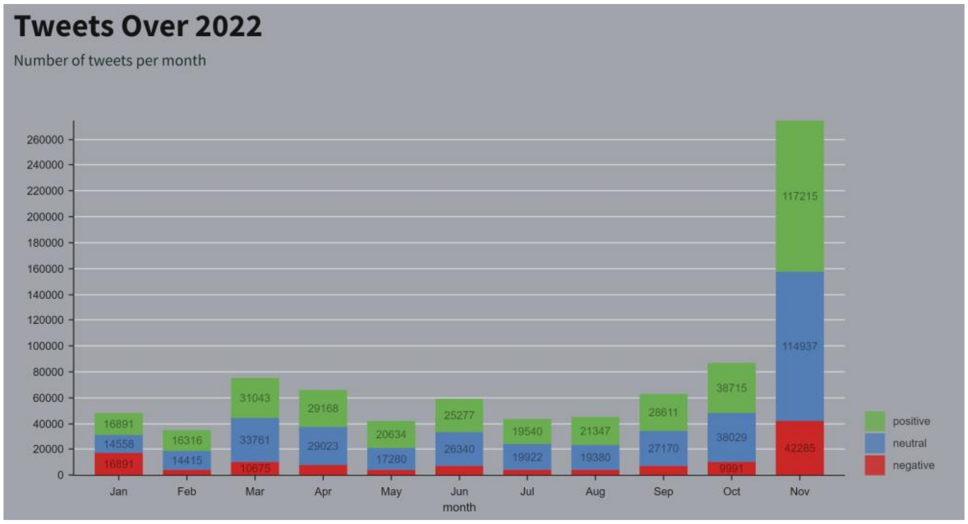


Fig 5: Plots a Bar Graph Months vs Number of Tweets

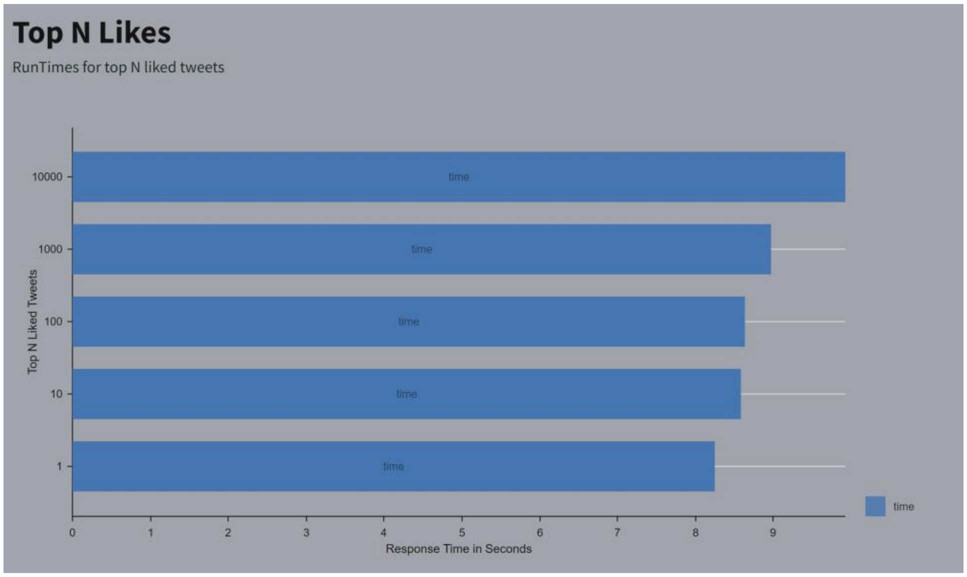
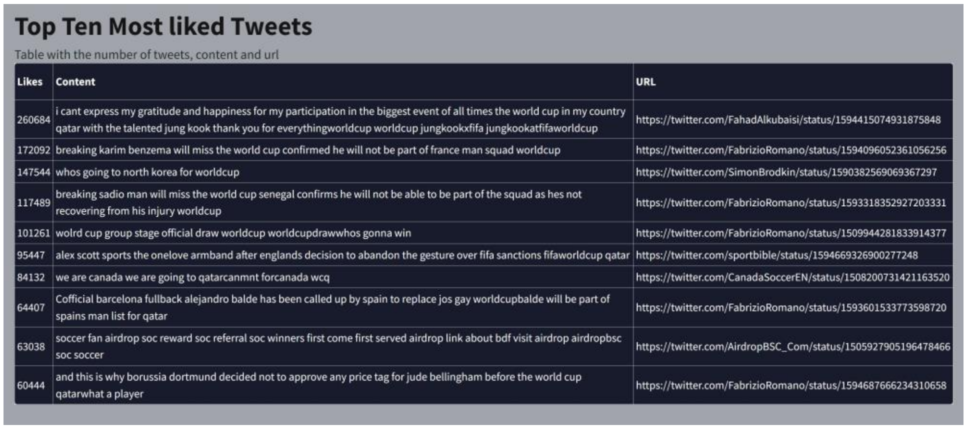


Fig 6: Plots a Bar Graph Query Response Time vs Top N Liked Tweets



Top Ten Most liked Tweets

Table with the number of tweets, content and url

Likes	Content	URL
260684	i cant express my gratitude and happiness for my participation in the biggest event of all times the world cup in my country qatar with the talented jung kook thank you for everythingworldcup worldcup jungkookxfifa jungkookatfifaworldcup	https://twitter.com/FahadAlkubaisi/status/1594415074931875848
172092	breaking karim benzema will miss the world cup confirmed he will not be part of france man squad worldcup	https://twitter.com/FabrizioRomano/status/1594096052361056256
147544	whos going to north korea for worldcup	https://twitter.com/SimonBrodkin/status/1590382569069367297
117489	breaking sadio man will miss the world cup senegal confirms he will not be able to be part of the squad as hes not recovering from his injury worldcup	https://twitter.com/FabrizioRomano/status/1593318352927203331
101261	wolrd cup group stage official draw worldcup worldcupdrawwhos gonna win	https://twitter.com/FabrizioRomano/status/1509944281833914377
95447	alex scott sports the onelove armband after englands decision to abandon the gesture over fifa sanctions fifaworldcup qatar	https://twitter.com/sportbible/status/1594669326900277248
84132	we are canada we are going to qatarcanmnt forcanada wcq	https://twitter.com/CanadaSoccerEN/status/1508200731421163520
64407	Cofficial barcelona fullback alejandro balde has been called up by spain to replace jos gay worldcupbalde will be part of spains man list for qatar	https://twitter.com/FabrizioRomano/status/1593601533773598720
63038	soccer fan airdrop soc reward soc referral soc winners first come first served airdrop link about bdf visit airdrop airdropbsc soc soccer	https://twitter.com/AirdropBSC_Com/status/1505927905196478466
60444	and this is why borussia dortmund decided not to approve any price tag for jude bellingham before the world cup qatarwhat a player	https://twitter.com/FabrizioRomano/status/1594687666234310658

Fig 7: A table representing the top ten liked tweets

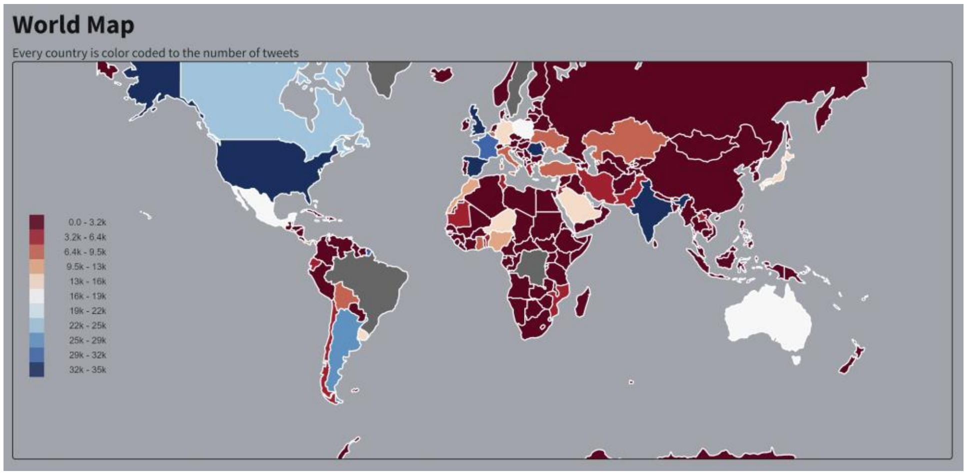


Fig 8: A choropleth world map with colors denoting number of tweets.

6. Queries Evaluation

For experimenting with queries runtime, we have run df.top(n, key) query for n=1, 10, 100, 1000, and 10000, and for key we have 3 parameters like count, reply count, and retweet count. Top is an action method of pyspark that uses in-memory space. So, for calculating query time on large data we have created Resilient Distributed Dataset of only necessary columns. The query time for the top 10 likes count was 8.5 seconds, for top 10 retweets count was 12 seconds and for top 10 reply count time was 9.19 seconds. Also, we have queried Top 1,10,100,10000 likes count. Result was top 1 = 8.2 while top 10000 = 9.9 seconds. Observation query time increases as the top count increases but not in significant amount. This was true for all the parameters like count, reply count and retweet count.

For scalable purposes we evaluated the top 10 most liked tweets on different sized dataset. On 60mb files it took 3 seconds to run. For 350mb file its runtime was 8.5 and for 3GB it was 115 seconds (about 2 minutes). Observation, it was required to remove unnecessary columns before calling top on rdd to make the query runnable.

7. Conclusion and Future work

The project consists of analysis of tweet sentiments and extracting general opinions of the people about FIFA world cup. The project employed NLTK to clean and process the data. Textblob was used to extract tweets polarity. 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. A visualization dashboard was built to help businesses manage customer demographic statistics. The dashboard tool was evaluated based on query response times to check for scalability.

Currently, we are limiting our dataset for sentiment analysis to the most common language, which is English, but using translating libraries can help us to expand this to multiple languages. In the future, Deep learning model could be implemented to improve the accuracy of sentiment analysis. Transformer models like BERT and ELMO which have proved to have state-of-the-art performances for Sentiment analysis could be fine-tuned for this application. A more interactive dashboard which has more query input fields depending on customer requirement could be built. Live tweet visualization could be the next feature to add to this dashboard tool which will visualize tweets based on specified timeframes.

8. Author Contribution

Vineet Dhaimodker - Gathering the data from the twitter, scraping and combining the data.

Vishv Patel - Setting up Hadoop and Spark on local machines. Data Pre-processing to convert raw data to efficient data.

Mihir Patel - Sentiment analysis and query processing. Assignment of sentiments using TextBlob library. Building the machine learning models and classifying tweets. Experimenting on different queries and evaluating results.

Yash Chinmay Gandhi - Created a Web Dashboard for data visualization with components like various bar charts, table and a world map.

8. References

[1] 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

[2] Lucas, G.M. et al. (2017) “GOAALLL!: Using sentiment in the World Cup to explore theories of emotion,” Image and Vision Computing, 65, pp. 58–65. Available at: https://doi.org/10.1016/j.imavis.2017.01.006.

[3] Yu, Y. and Wang, X. (2015) “World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans’ tweets,” Computers in Human Behavior, 48, pp. 392–400. Available at: https://doi.org/10.1016/j.chb.2015.01.075.

[4] Fan, M. et al. (2019) “Twitter-based BIRGing: Big Data Analysis of English national team fans during the 2018 FIFA World Cup,” Communication &amp; Sport, 8(3), pp. 317–345. Available at: https://doi.org/10.1177/2167479519834348.

[5] Camilleri, L. (2019). Natural language processing for sentiment analysis (Bachelor's dissertation).

[6] Khader, M., Awajan, A. and Al-Naymat, G. (2018) “The effects of natural language processing on Big Data Analysis: Sentiment Analysis Case Study,” 2018 International Arab Conference on Information Technology (ACIT) [Preprint]. Available at: https://doi.org/10.1109/acit.2018.8672697.

[7] Kanakaraj, M. and Guddeti, R.M. (2015) “NLP based sentiment analysis on Twitter data using ensemble classifiers,” 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN) [Preprint]. Available at: https://doi.org/10.1109/icscn.2015.7219856.

[8] Taboada, M. et al. (2011) “Lexicon-based methods for sentiment analysis,” Computational Linguistics, 37(2), pp. 267–307. Available at: https://doi.org/10.1162/coli_a_00049.

[9] Shayaa, S. et al. (2018) “Sentiment analysis of Big Data: Methods, applications, and open challenges,” IEEE Access, 6, pp. 37807–37827. Available at: https://doi.org/10.1109/access.2018.2851311.

[10] Medhat, W., Hassan, A. and Korashy, H. (2014) “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, 5(4), pp. 1093–1113. Available at: https://doi.org/10.1016/j.asej.2014.04.011.

[11] Gautam, G. and Yadav, D. (2014) “Sentiment analysis of Twitter data using machine learning approaches and semantic analysis,” 2014 Seventh International Conference on Contemporary Computing (IC3) [Preprint]. Available at: https://doi.org/10.1109/ic3.2014.6897213.

[12] Alsaeedi, A. and Zubair, M. (2019) “A study on sentiment analysis techniques of Twitter data,” International Journal of Advanced Computer Science and Applications, 10(2). Available at: https://doi.org/10.14569/ijacsa.2019.0100248.

[13] Gaglio, S., Lo Re, G. and Morana, M. (2016) “A framework for real-time Twitter data analysis,” Computer Communications, 73, pp. 236–242. Available at: https://doi.org/10.1016/j.comcom.2015.09.021.

[14] Stojanovski D, Dimitrovski I, Madjarov G. Tweetviz: Twitter data visualization. Proceedings of the data mining and data warehouses. 2014 Oct;1(2).

[15] Yaqub, U. et al. (2018) “Analysis and visualization of subjectivity and polarity of Twitter location data,” Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age [Preprint]. Available at: https://doi.org/10.1145/3209281.3209313.

[16] Gilmour, J.B., Lui, A.W. and Briggs, D.C. (1986) EMR, Amazon. Amazon. Available at: https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/ (Accessed: December 6, 2022).

[17] Big data (2015) Amazon. Springer, India, Private Ltd. Available at: https://aws.amazon.com/big-data/what-is-spark/ (Accessed: December 6, 2022).

[18] (no date) Home. Available at: https://nivo.rocks/ (Accessed: December 6, 2022).

[19] Material-ui: A popular react UI framework (no date) UI. Available at: https://v4.mui.com/ (Accessed: December 6, 2022).

[20] Harrison, P. (2021) React table: A complete tutorial with examples, LogRocket Blog. Available at: https://blog.logrocket.com/complete-guide-building-smart-data-table-react/ (Accessed: December 6, 2022).

[21] *ML tuning: Model selection and hyperparameter tuning* (no date) *ML Tuning - Spark 3.3.1 Documentation*. Available at: https://spark.apache.org/docs/latest/ml-tuning.html (Accessed: December 6, 2022).