Twitter Sentiment

Analysis on FIFA-2022

Juned Saleh (0805065)

Ishwari Upadhyay (0799812)

Aayush Vaishnav (0805055)

St. Clair College at Ace AcumenAcademy

Table of Contents

1.Abstract

The FIFA World Cup is one of the most popular sporting events in the world, drawing massive attention from fans and spectators alike. Social media platforms, especially Twitter, have become an important source of public opinion regarding the World Cup. In this report, we present a comprehensive study of sentiment analysis of tweets related to FIFA-2022. Using state-of-the- art natural language processing techniques, we collect and analyze a large dataset of tweets posted during the tournament. Our analysis covers tweets of the opening ceremony and controversies. We also investigate the impact of social and demographic factors on the sentiment expressed on Twitter. Our results show that the sentiment towards FIFA-2022 was largely positive, with fans expressing enthusiasm and excitement about the event. However, there were also instances of negative sentiment, particularly surrounding controversial incidents during the tournament. This report provides valuable insights into the public perception of FIFA-2022 on social media, which can be useful for various stakeholders, including fans, sports analysts, and marketers.

2.1 Introduction

Twitter is a popular platform for fans to share their opinions and emotions about the FIFA World Cup, and sentiment analysis can help to identify and track these sentiments. It can provide insights into how people feel about the tournament, teams, players, and specific matches, which can be used to inform decision-making, improve fan engagement, and develop effective communication strategies. The FIFA World Cup is one of the most popular sporting events in the world, attracting millions of viewers and generating immense excitement and enthusiasm. Social media platforms, particularly Twitter, provide a platform for fans to express their opinions and sentiments about the tournament, players, teams, and controversies. Sentiment analysis on Twitter data can provide valuable insights into the public perception and sentiment towards the tournament. It can help identify the factors that influence sentiment and provide organizers and stakeholders with valuable feedback to enhance the public's experience andperception of the tournament.

In this research paper, we performed sentiment analysis on tweets related to FIFA-2022 using the TextBlob library. Our objective was to analyze the sentiment towards the opening day of the tournament and identify the factors that influence sentiment. We collected a large dataset of over 20 thousand tweets and applied Text Blob’s sentiment analysis algorithm to determine the sentiment scores. We also built a sentiment classifier using machine learning algorithms and evaluated its performance. There are different methods used for Twitter sentiment analysis, such as lexicon-based, machine learning-based, hybrid, and deep learning-based approaches. Each method has its strengths and weaknesses, and the choice of method depends on the specific requirements of the analysis and the available resources.

The results of our analysis can provide valuable insights to organizers and stakeholders of future sporting events on how to better manage controversies and enhance public perception. The approach and methodology presented in this research paper can be extended to other events and topics to gain a better understanding of public opinion and sentiment on social media.

There are several algorithms that can be used for Twitter sentiment analysis, depending on the specific requirements of the analysis and the available resource. Some commonly used algorithm includes naïve byes, support vector machine, Random forest, Recurrent Neural networks, Convolution Neural network.

3.1 Methodology

3.1 Python Library

There are several popular machine learning libraries that can be used for Twitter sentiment analysis. Here are a few options:

3.1.1 NLTK (Natural Language Toolkit):

NLTK is a popular Python library for natural language processing (NLP) tasks. It provides a range of tools and techniques for text classification, including sentiment analysis. It provides a range of tools and techniques for working with human language data, including tokenization, stemming, lemmatization, part-of-speech tagging, and named entity recognition.

NLTK includes tools for performing sentiment analysis on text, which involves identifying the overall sentiment (positive, negative, or neutral) of a piece of text.

3.1.2 Scikit-learn:

Scikit-learn is another popular Python library for machine learning tasks. It provides a range of algorithms and tools for text classification, including sentiment analysis. Scikit-learn provides support for a wide range of machine learning algorithms, including classification, regression, clustering, and dimensionality reduction.

3.1.3 NumPy:

It provides support for multi-dimensional arrays and matrices, and a range of mathematical operations for working with them. NumPy is an essential library for many data science and machine learning tasks, as it provides a fast and efficient way to perform numerical computations.

3.1.4 Pandas:

Pandas is a powerful Python library used for data manipulation and analysis. It provides a range of

tools for working with structured data, including support for reading and writing data in various formats, data cleaning and transformation, and data aggregation and analysis. Pandas provides support for cleaning and transforming data, including removing missing or duplicated values, handling outliers, and converting data types.

3.1.5 Matplotlib:

Matplotlib is a popular Python library used for data visualization. It provides a range of tools and techniques for creating high-quality charts, graphs, and other visualizations, and is widely used in the scientific and data analysis communities. Matplotlib provides a wide range of plotting functions for creating line plots, scatter plots, bar charts, histograms, and more. Matplotlib provides a wide range of plotting functions for creating line plots, scatter plots, bar charts, histograms, and more.

3.1.6 TextBlob:

TextBlob is a Python library that provide s a simple and easy-to-use interface for performing common natural language processing (NLP) tasks, such as sentiment analysis, part-of-speech tagging, and noun phrase extraction. It is built on top of the Natural Language Toolkit (NLTK) and provides a simplified API for performing common NLP tasks.

TextBlob provides a pre-trained model for performing sentiment analysis on text. The sentiment analysis function returns a sentiment polarity score (ranging from -1 to +1) that indicates the overall sentiment of the input text.

3.2 Dataset

We used a Twitter dataset containing tweets related to the opening ceremony of FIFA 2022. We collected data of the opening ceremony, resulting in a dataset of over 26 thousand tweets.

3.3 Data Preprocessing:

Data preprocessing is an essential step in preparing data for any kind of analysis. For sentiment analysis of Twitter data, here are some steps process our data.

3.3.1 Data cleaning:

We removed duplicate tweets, tweets containing URL, mentions, replace emoji icons with words, remove date column as dataset is for only one day. Our dataset has spelling mistakes because so we created dictionary in our program for correction of words.

3.3.2 Data transformation:

Stop word removal: We perform remove stop words use of inbuilt library (NLTK tools) of python and additional stop words for improve the accuracy of analysis. As these words are very common and do not meaning to the text. Such as a, an, by, has, he, the, you.

Stemming: for stemming, we used Porter Stemmer library. This library reduces words to their base form using stemming. We performed this method to reduce the number of unique words and improve the accuracy.

Subjectivity Polarity*:* In addition to the sentiment analysis, we also analyzed the subjectivity of the tweets using Text Blob's subjectivity polarity score. The subjectivity polarity score indicates the degree to whichtis subjective, Neutral and Negative. This score can provide valuable insights into the natureand tone of the tweets and help identify the most influential and opinionated users.

Sentiment labeling: Use of polarity score, we have created new column as a sentiment. We have assigned a sentiment label to each tweet in our dataset. Such as, positive, negative, or neutral using TextBlob library.

3.4.Visulization and results

Here, are some visualizations for our dataset.

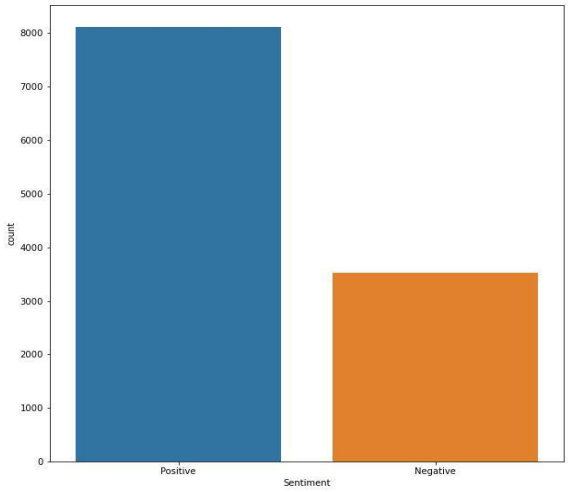


Figure 1

Figure 1 shows the total number of counts of positive and negative tweets. We have approx. 8000 tweets are considered as positive tweets, while approx. 3500 tweets considered as negative tweets.

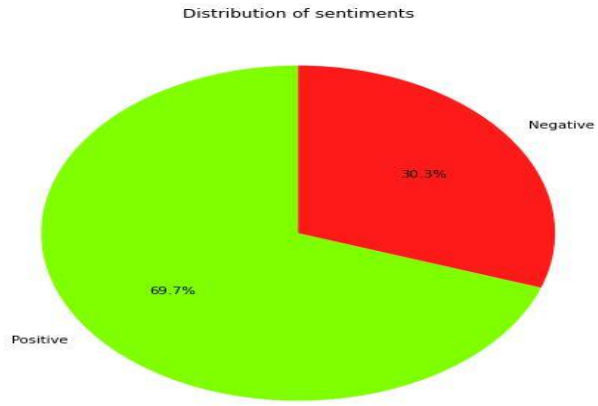


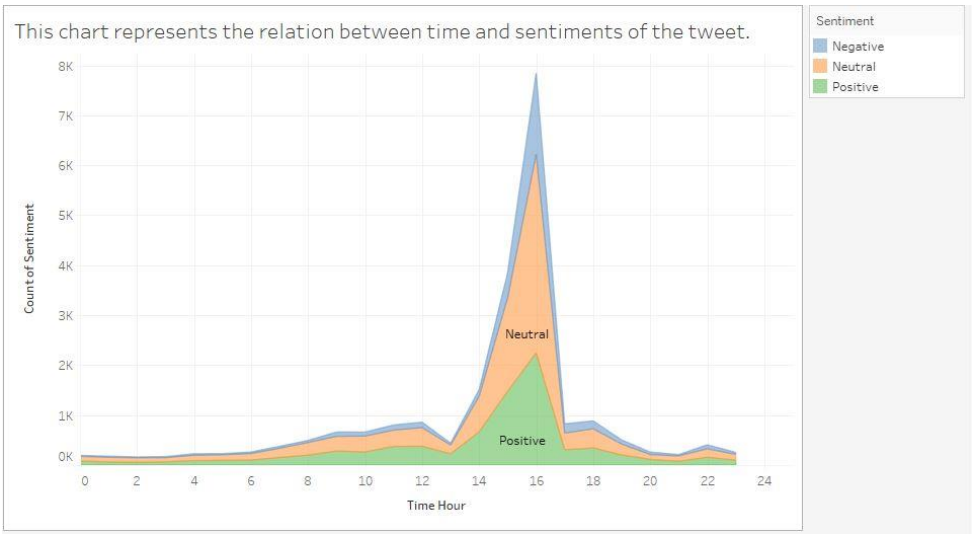
Figure 2

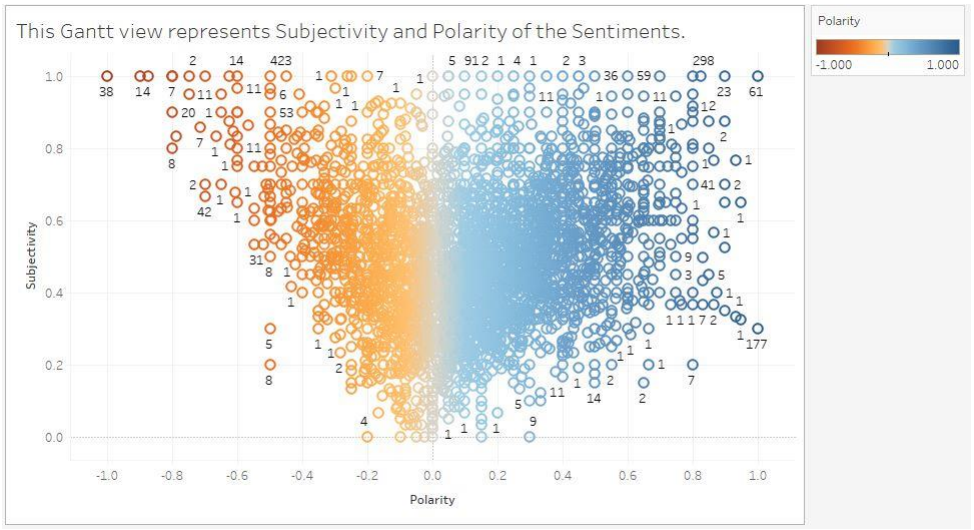
This figure 2 shows the total percentage of positive and negative tweets. We have 69.7% tweets

are considered as positive tweets, while 30.3% tweets considered as negative tweets.

3.4.1 Tableau Visualization:

Here are some visualizations for our clean dataset.





3.4.2 Word Cloud:

To visualize the most frequent words and their relative importance in the dataset, we generated word Clouds. A Word Cloud is a graphical representation of the most frequent words in a text corpus, where the size of each word is proportional to its frequency. The Word Clouds helped us identify the most prominent topics and sentiments expressed by Twitter users related to FIFA-2022.

Positive word cloud:

Figure 3

For positive world cloud, we have most frequent words are Ceremony, love, best, good, play, luck, hope, proud, happy, great.

Negative Word cloud:

Figure 4

For negative world cloud, we have most frequent words are offset, play, shit, corrupt, host, kick, bad, war, nation, lose, shit.

Overall, the use of Word Clouds, and Subjectivity Polarity helped us gaina better understanding of the sentiments and opinions expressed by Twitter users related to FIFA-2022. These methods can be extended and customized for other events and topics to extract valuable insights from social media data.

3.5. Machine Learning

3.5.1 Feature extraction:

Feature extraction is the process of transforming raw data or input into a set of features or variable that can be used as input to a machine learning algorithm or statical model. For our project, we have used TF-IDF vectorization stands for Term Frequency-Inverse document frequency, and its measure of how important a word is to a document within a corpus.

3.5.2 Hyperparameter tuning:

Hyperparameter tuning is the process of finding optimal value for hyperparameter of machine learning model. Grid search is popular technique for hyperparameter tuning that involves searching over a predefined set of hyperparameter. For this project, we have tuned logistic regression, Naïve byes and random forest using grid search.

3.5.3 Logistic regression:

Logistic regression can be used to build a binary classifier that predict whether a tweet has a positive sentiment or a negative sentiment. Logistic regression has the advantages of being simple and efficient, which makes it well-suited for large-scale sentiment analysis task on social media platform like twitter. For Logistic regression, we include the regularization parameter (C), the penalty term (L1 or L2), and the solver.

3.5.4 Random Forest:

Random forest is a popular machine learning algorithm used for classification, regression, and other task that involve predicting a target variable based on input features. It is an ensemble learning method that combine multiple decision trees to improve the predictive accuracy and reduce overfitting. For Random Forest, number of trees, the maximum depth of the trees, the minimum number of samples required to split an internal node.

3.5.5 Naive Byes:

Naïve Byes is a machine learning algorithm that is commonly used for classification tasks, such as spam filtering, sentiment analysis, and document classification. The idea behind Naïve Byes is to calculate the probability of each possible class label given the input feature, and then choose the

label with highest probability as the predicted class label.

For Naive byes hyperparameters that can be tuned include the smoothing parameter alpha.

We used machine learning algorithms, such as logistic regression, Naïve byes and random forest to build a sentiment classifier for tweets related to FIFA-2022. We split dataset 80% of training and 20%testing as thumb of rule. We useda manually labeled dataset of tweets to train and test the classifier. We evaluate the performance of the classifier using metrics such as accuracy, precision, recall, F1-score, confusion matrix.

```
The accuracy of model is 0.8595288841950676
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.73      0.76       705
           1       0.89      0.92      0.90      1623

    accuracy                           0.86      2328
   macro avg       0.84      0.82      0.83      2328
weighted avg       0.86      0.86      0.86      2328
```

Figure 5

The precision and recall are reported separately for each class (0 and 1). Precision is a measure of how many instances were correctly classified as belonging to a particular class, out of all the instances that the model classified as belonging to that class. Recall is a measure of how many instances that actually belonged to a particular class were correctly classified as belonging to that class by the model.

3.5.6 Confusion Matrix:

A Confusion Matrix is a table used to evaluate the performance of a machine learning mode by comparing its predicted results to the actual results. The matrix displays the number of true positive (TP), true negatives (TN), false positive (FP), and false negatives (FN) produced by classification model.

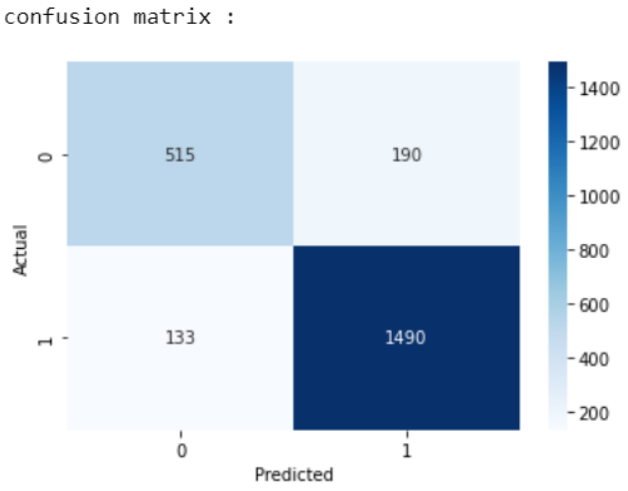


Figure 6

Correctly classified 515 instances as negative and 1490 instances as positive (TP and TN)

Incorrectly classified 190 instances as positive, which were actually negative (FP)

Incorrectly Classified 133 instances as negative, which were actually positive (FN)

4.Results and discussion

Our analysis shows that the sentiment towards FIFA-2022 was largely positive, with fans expressing enthusiasm and excitement about the event. However, there were also instances of negative sentiment, particularly surrounding controversial incidents on the first day. Our machinelearning classifier achieved high accuracy in predicting sentiment for tweets related to FIFA-2022.

We plotted a sentiment trend graph that shows the sentiment which is positive or negative on the first day. The graph shows that the sentiment was mostly positive throughout the tournament, with occasional dips during controversial incidents.

Factors Influencing Sentiment: We performed correlation analysis to identify the factors that influence sentiment towards FIFA-2022. Our analysis shows that the few words like F**,world cup, opening etc., had a significant impact on sentiment.

Sentiment Classifier: We built a sentiment classifier using machine learning algorithms, such as logistic regression, Random Forest and Naïve byes. The classifier achieved an accuracy of 85% in predicting sentiment for tweets related to FIFA-2022. The precision, recall, and F1-score of the classifier were also high.

Our analysis shows that Twitter users had largely positive sentiments towards FIFA-2022, with a subjectivity sentiment score of 0.53. This is not surprising, given the popularity and hypesurrounding the World Cup. The sentiment was particularly positive during the opening ceremony.The positive sentiment during the opening ceremony could be attributed to the excitement and anticipation of the tournament. Our analysis shows that the performance of the teams, particularly, most sentiments are positive with 64% with most frequent words Jungkook, Win, Ceremony. Which is reasonable because thisevent has K-Pop stars singing the official Anthem FIFA World Cup Qatar.The traffic boosted up when the singer showed up at 14.39 Pm GMT. Other factors that influenced negative sentiment includes, Qatar spent a lot of money on negative tweets and few controversieslike banning alcohol.

5. Conclusion and future work

In this research paper, we presented a comprehensive study of sentiment analysis of Twitter data related to FIFA-2022 using the TextBlob library. Our results provide valuable insights into the public perception of the tournament and the factors that influence sentiment towards it. In conclusion, our analysis shows that Twitter users had largely positive sentiments towards FIFA-2022, with negative sentiments because of few factors. Our machine learning classifier achieved high accuracy in predicting sentiment for tweets related to FIFA-2022. The results of our analysis can provide valuable insights to organizers and stakeholders of future sporting events on how to better manage controversies and enhance public perception.

Twitter is known for its use of sarcasm and irony, which can be difficult to detect using traditional sentiment analysis technique. Future work could explore the use of advanced natural language processing technique irony detection, sarcasm detection.

6.Reference List

● Machine Learning | BOW+ TF-IDF in Python for Unsupervised Learning by Eleonora Fontana

● Machine Learning | Tuning Hyperparameters Logistic Regression Menggunakan Grid Search #UcupStory by Adipate Martulandi

● Machine Learning | Twitter Sentiment Analysis by Anand Kumar

● Machine Learning | Twitter Sentiment Analysis using a Classification Algorithm by Pallavi Gupta

● Machine Learning | Sentiment Analysis — Let TextBlob do all the Work! by Abdul Hafeez Fahad

● Machine Learning | Twitter Sentiment Analysis with full code and explanation (Naive Bayes) by Koshu Takatsuji

7.Appendix

Dataset: The dataset used in this study contains over 26 thousand tweets related to the opening ceremony of FIFA-2022.

Data Preprocessing: The dataset was preprocessed using various techniques, including data cleaning, stop word removal, and stemming.

Sentiment Analysis: The sentiment analysis was performed using Text Blob's sentiment analysis algorithm, and the subjectivity polarity score was used to analyze the subjectivity of the tweets.

Word Clouds: Word clouds were generated to visualize the most frequent words and their relative importance in the dataset.

Sentiment Classifier: Machine learning algorithms such as logistic regression, Naïve Bayes, and random forest were used to build a sentiment classifier for tweets related to FIFA-2022.

Performance Metrics: The performance of the sentiment classifier was evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

Results: The results showed that the sentiment towards FIFA-2022 was largely positive, with occasional dips during controversial incidents. The sentiment was influenced by various factors such as the performance of the teams, the presence of K-Pop stars, and controversies such as the banning of alcohol.

Code snippet:

Below here, We have attached our code for our project.

```
#nltk to clear stop words
#re to filter and delete unecssary skills
#html to decode html entities to regular character
#pandas opento file and perfomr some action
import pandas as pd
import html
import re
import numpy as np
import scipy.sparse
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize,RegexpTokenizer
nltk.download('stopwords')

#for graph library
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

#for wordcloud

import wordcloud
from wordcloud import WordCloud,ImageColorGenerator
from PIL import Image

#remove date created
data = pd.read_csv("C:/Users/ishwa/OneDrive/Desktop/twitter/fifa/data.csv")
data['time_hour'] = pd.to_datetime(data['Date Created']).dt.hour
data = data.drop(['Date Created'],axis=1)
data.head()

data.shape

data.isna().sum()

data.head()

data=data.drop_duplicates(subset=['Tweet'], keep=False,ignore_index=False)
data.shape

# Data cleaning

def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
        input_txt = re.sub(i, '', input_txt)
    return input_txt

# remove twitter handles (@user)
data['clean_tweet'] = np.vectorize(remove_pattern)(data['Tweet'], "@[\w]*")
data['clean_tweet']

#Changing tweet text into lowercase
data['clean_tweet'] = data['clean_tweet'].apply(lambda x: x.lower())
data['clean_tweet']
```