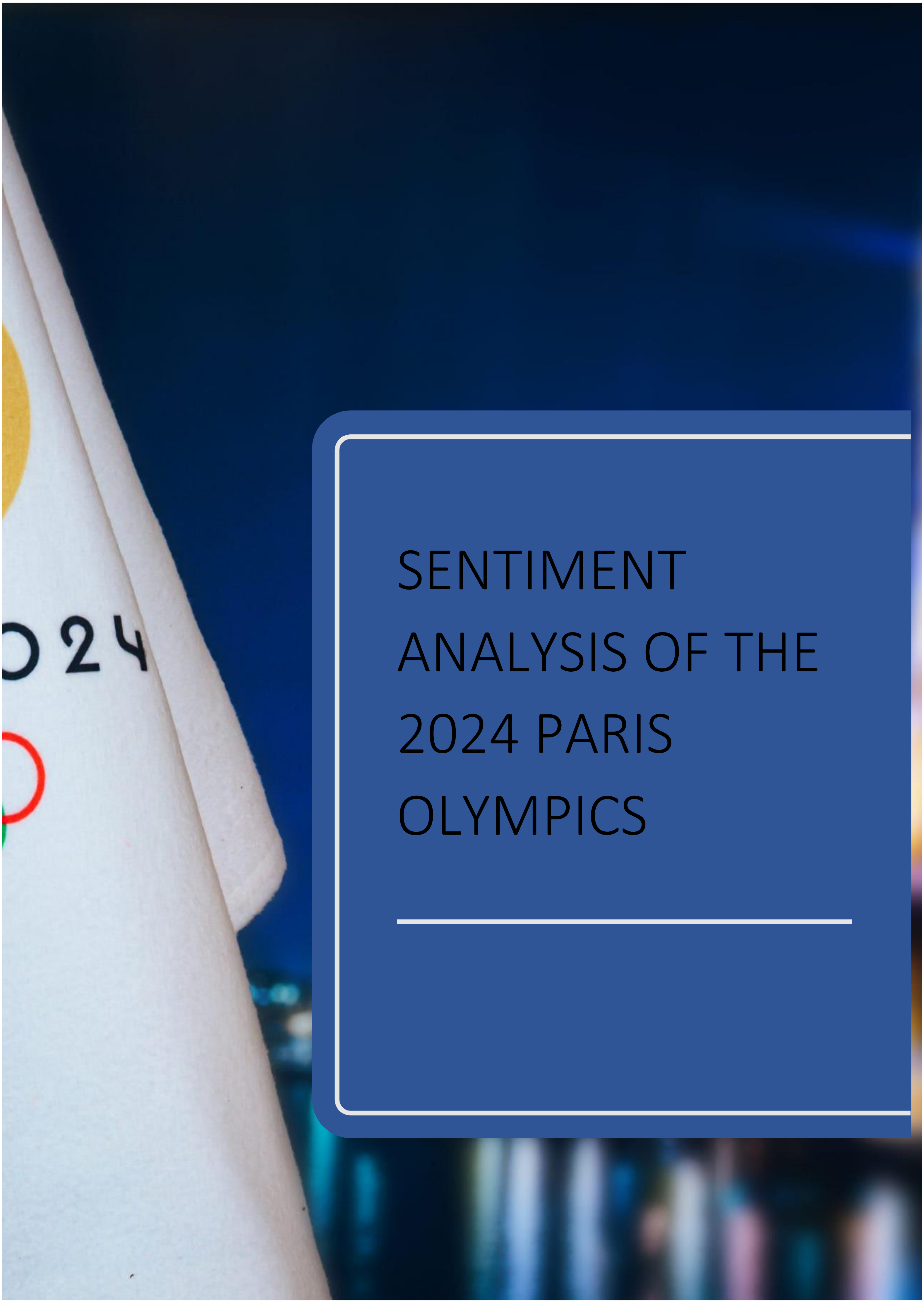


SENTIMENT ANALYSIS OF THE 2024 PARIS OLYMPICS

TABLE OF CONTENTS

BUSINESS UNDERSTANDING

Overview

The Paris Olympics is a global sporting event that has garnered significant attention and engagement across various social media platforms. Analyzing public sentiment regarding the Olympics can provide valuable insights into how athletes, countries and the overall event are perceived. This analysis can benefit sports organizations, media outlets, sponsors offering feedback on public perception, performance and engagement levels thus helping to tailor content and marketing strategies. Sentiment analysis can also benefit city officials to improve planning and address concerns such as health and sanitation.

The goal of this project is to perform a comprehensive sentiment analysis of social media content related to this year's Paris Olympics to understand public sentiment, identify emerging trends and provide a comprehensive understanding of how different aspects of the Olympics resonate with audiences worldwide.

Challenges

1. Social media data is noisy and unstructured presenting challenges for accurate analysis.
2. Distinguishing between positive, negative and neutral sentiments can be difficult especially when dealing with multilingual content thus affecting sentiment analysis accuracy
3. The volume of social media posts and comments can be overwhelming particularly during major events like the Olympics. Managing and processing large volumes of real-time data necessitates efficient data handling and processing techniques.
4. Interpreting context and sarcasm an extra layer of complexity as the sentiment expressed may not always align with the literal meaning of the words used. Social media content often includes informal language, slang and nuanced expressions that can skew sentiment analysis.

Proposed Solutions

1. Use API access to collect data from major social media platforms and ensure compliance with platform policies and data protection regulations.
2. Implement text normalization, tokenization and content filtering while utilizing language detection and translation tools for multilingual data handling.
3. Employ advanced natural language processing models like BERT or GPT for sentiment classification incorporating sarcasm detection and contextual analysis for improved accuracy.
4. Create an interactive dashboard using Tableau to display sentiment trends and insights with features for data filtering and exploring different aspects of the data.

Stakeholders

1. Organizers of the Paris Olympics 2024 - Sentiment analysis helps them gauge public opinion allowing them to make informed decisions and adjust their strategies accordingly.
2. Sponsors - Sentiment analysis helps them understand how their brand is perceived in relation to the Olympics.
3. Media outlets - Sentiment analysis provides them with insights into public interest and trending topics.
4. Fans and general public - They are the primary audience for the Olympics and their sentiment directly impacts the event's success.
5. Athletes - They are the central figures of the Olympics and public sentiment towards them can affect their performance and well-being.
6. Local authorities and businesses in Paris - The Olympics significantly impact the host city and sentiment analysis can help gauge public opinion on local issues related to the event.

Success Metrics

1. Accuracy – The proportion of correctly classified sentiments (positive, negative, neutral) out of all sentiments predicted by the model.
   85% - 90%

2. Precision - The proportion of true positive sentiment predictions (correctly identified positive tweets) out of all predicted positives.
   80% - 90% for both positive and negative sentiment classes.
   75% - 85% for the neutral class.

3. Recall - The proportion of true positive sentiment predictions out of all actual positives.
   75% - 80% for all sentiment classes.

4. F1 Score - The harmonic mean of Precision and Recall that provides a single metric that balances both precision and recall.
   0.75 to 0.85

Conclusion

This sentiment analysis project aims to deliver a comprehensive understanding of public opinion about the Paris Olympics by leveraging social media data. By addressing the challenges of data quality, sentiment accuracy, multilingual content and implementing advanced NLP techniques, the project will provide actionable insights to the aforementioned stakeholders. Successful execution will enable better engagement strategies and enhance the overall experience of the Olympics for audiences worldwide

Problem Statement

The Paris Olympics is a high-profile event that generates a substantial volume of unstructured social media data that reflects public sentiment. The challenge lies in effectively analyzing this vast and diverse stream of data while also tackling challenges such as language differences, sentiment variations and contextual meanings in order to provide accurate and actionable insights.

Objectives

Main Objective

Develop a comprehensive social media sentiment analysis model that accurately captures and interprets public sentiment about the Paris Olympics from social media data.

Specific Objectives

1. To extract, preprocess and clean social media data from multiple platforms addressing quality issues and handling multilingual content related to the Paris Olympics.
2. To develop and train advanced natural language processing models to accurately classify sentiments incorporating techniques to handle sarcasm and contextual nuances.
3. To create interactive visualizations to display sentiment trends and key events providing actionable insights to stakeholders based on comprehensive analysis of public opinions.

DATA UNDERSTANDING

Data Sources

We extracted data from X using Octorparse Webscraping Tool.The focus was on posts mentioning Paris Olympics and relevant hashtags.

Datasets

Tweets in the form of hashtags, comments and retweets discussing the various aspects of the Paris Olympics.

Relevance of The Data

The data sources and datasets identified for this project are highly relevant to analyzing public sentiment surrounding the Paris Olympics. A Social media platform like X (formerly Twitter) captures immediate reactions, discussions and emotional responses from a global audience thus providing a rich source of unfiltered public sentiment. Relevant hashtags allow for more targeted analysis potentially revealing topic-specific sentiments.

DATA PREPARATION

The data processing involved steps to analyze and clean a dataset of tweets related to the 2024 Paris Olympics. The initial dataset was composed of multiple CSV files but was later merged into a single DataFrame. A 'DataUnderstanding' class was created to check the dataset's shape, columns, unique values, missing data and duplicates.

Upon examining the dataset we noted that some columns had missing values. The dataset also contained a large number of duplicated records which were primarily false positives due to pa rtial similarities rather than exact matches across all columns.

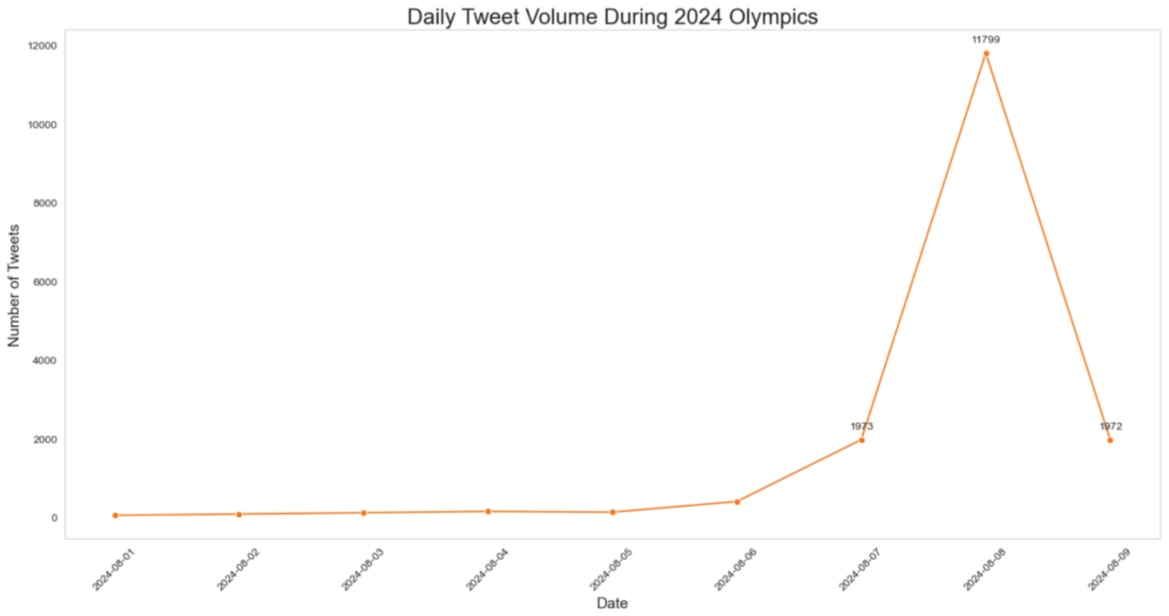
While the dataset contained many apparent duplicates, the true duplicates were fewer. The analysis revealed discrepancies and potential errors in data identification prompting further data cleaning to ensure accurate results.

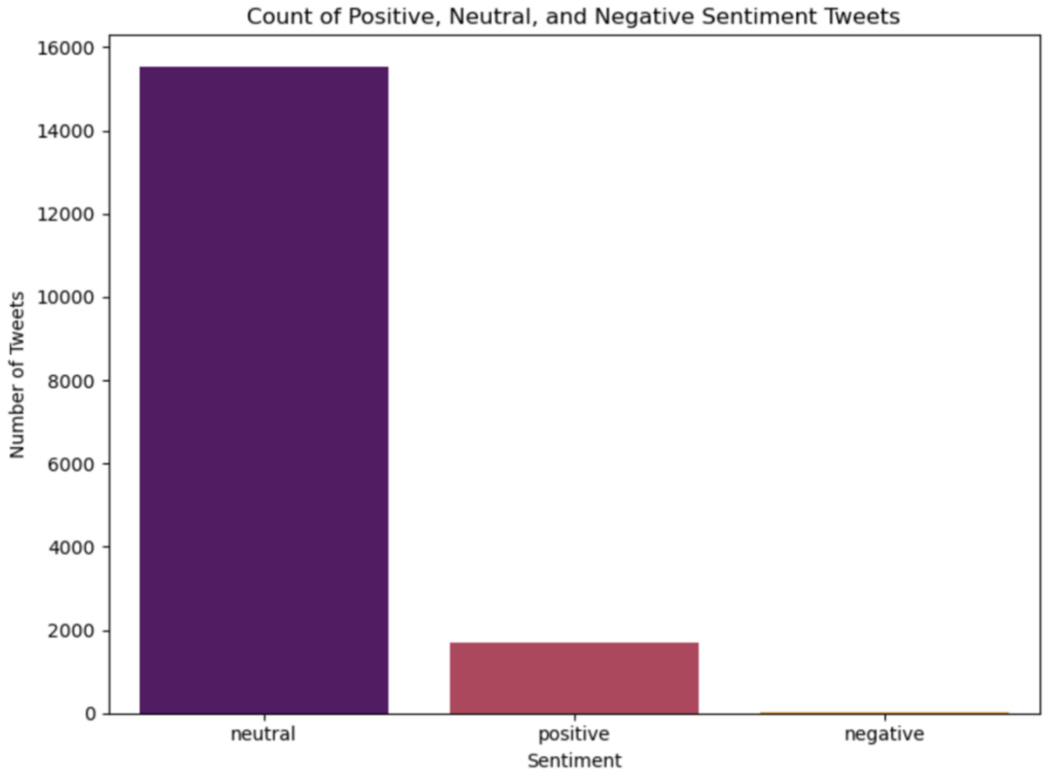
Data Cleaning

In the data cleaning process, irrelevant columns such as 'Tweet_Image_URL', 'Web_Page_URL' and 'Tweet_AD' were dropped, duplicates and rows with null values were also removed. Column names were stripped of whitespace and the 'Tweet_Timestamp' column was converted to datetime format. The cleaned dataset was then filtered to include only data from the year 2024. Data completeness was ensured by confirming the absence of null values and addressing consistency issues where rows were incorrectly identified as duplicates. Engagement columns were converted to numeric values and text preprocessing involved removing URLs (unlimited resource locators), mentions, hashtags, punctuation and numbers while retaining relevant tokens.

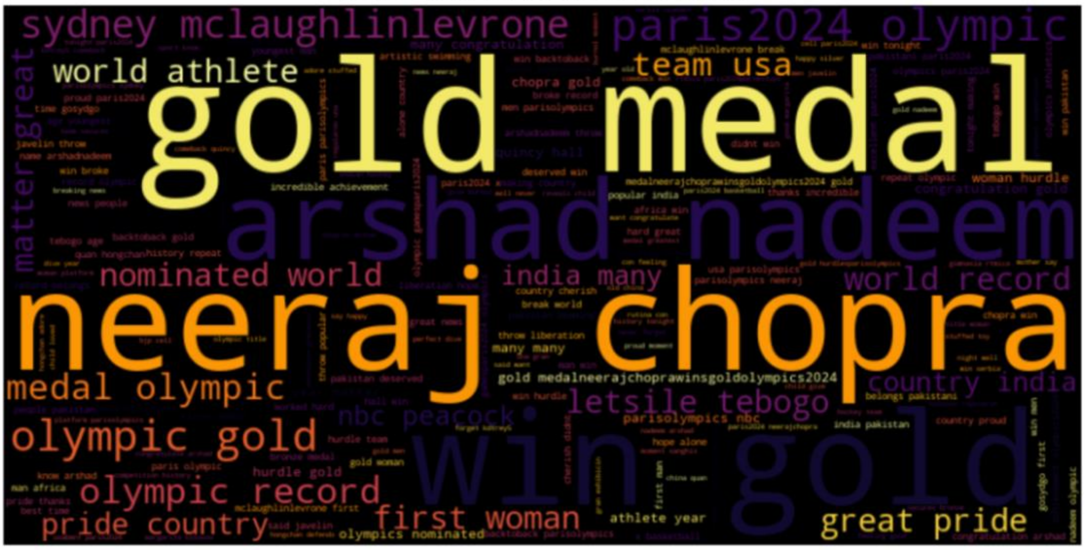
Exploratory Data Analysis

Exploratory Data Analysis involved examining and visualizing data to understand its main characteristics, patterns and relationships. It revealed trends, correlations and anomalies in the tweet sentiments and engagement around the Olympics which may not have been immediately apparent in the raw data. We made use of bar plots, heat maps, word clouds and line plots to visualize the Paris Olympics 2024.

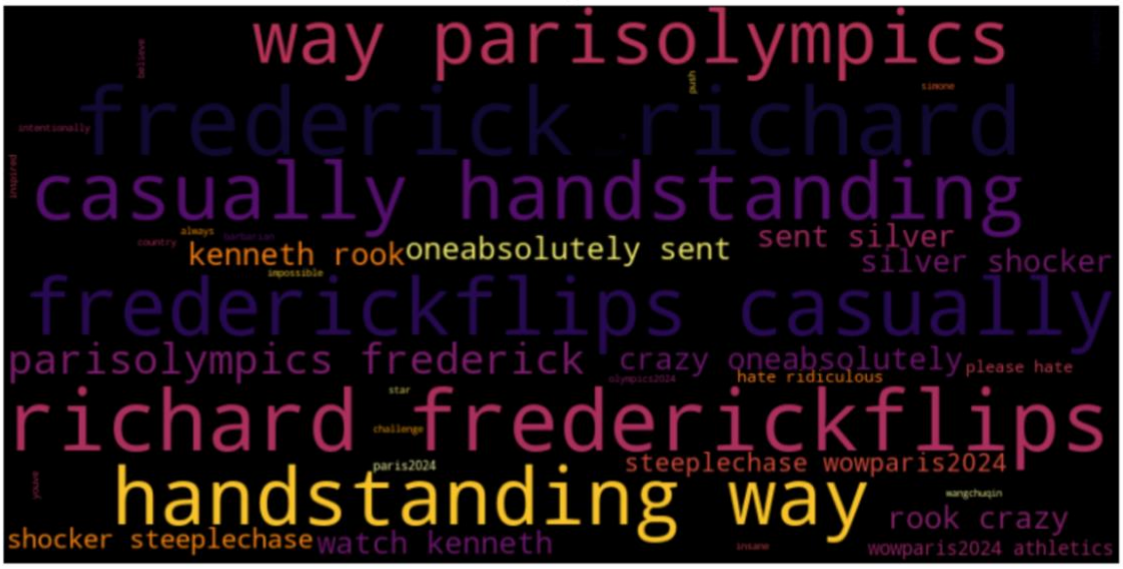

Daily Tweet Volume During 2024 Olympics
Number of Tweets
0
2000
4000
6000
8000
10000
12000
11799
1973
1972
2024-08-01
2024-08-02
2024-08-03
2024-08-04
2024-08-05
2024-08-06
2024-08-07
2024-08-08
2024-08-09
Date

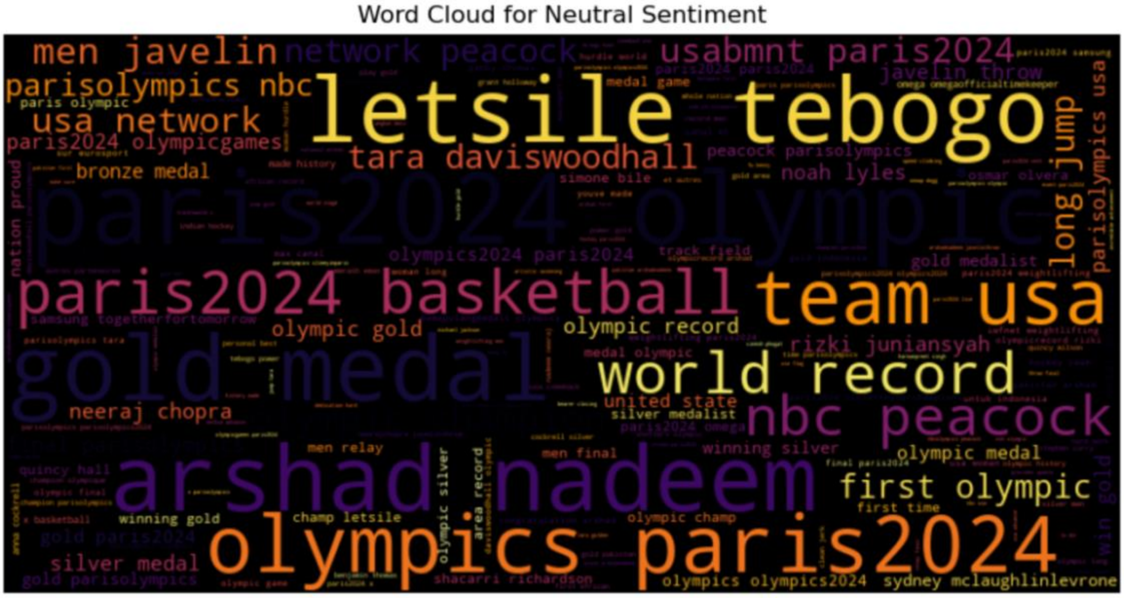

Count of Positive, Neutral, and Negative Sentiment Tweets
Number of Tweets
0
2000
4000
6000
8000
10000
12000
14000
16000
neutral
positive
negative
Sentiment

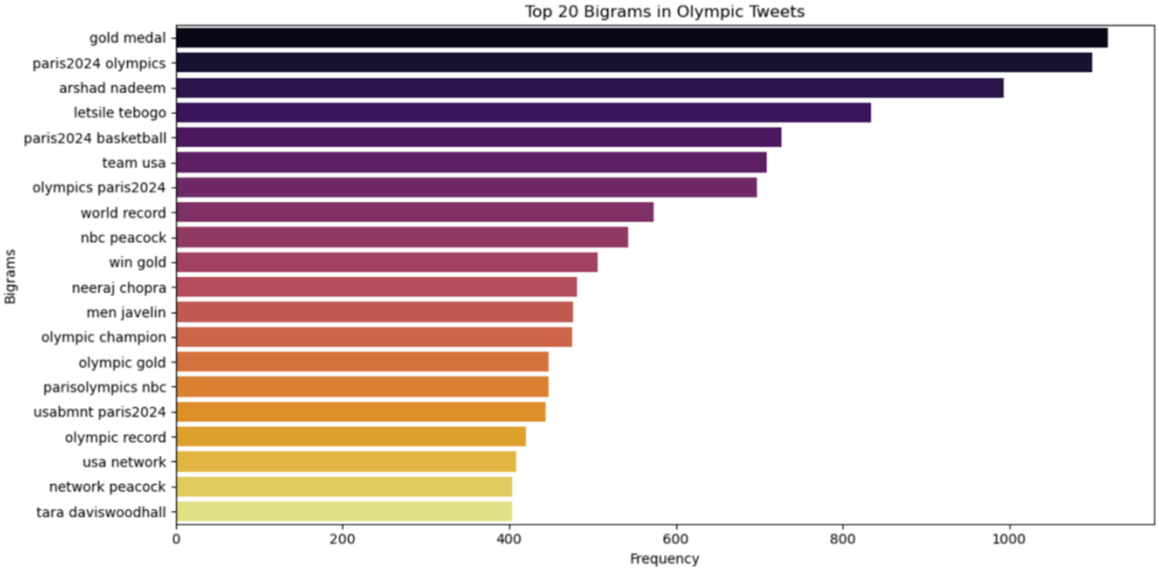
Word Cloud for Positive Sentiment

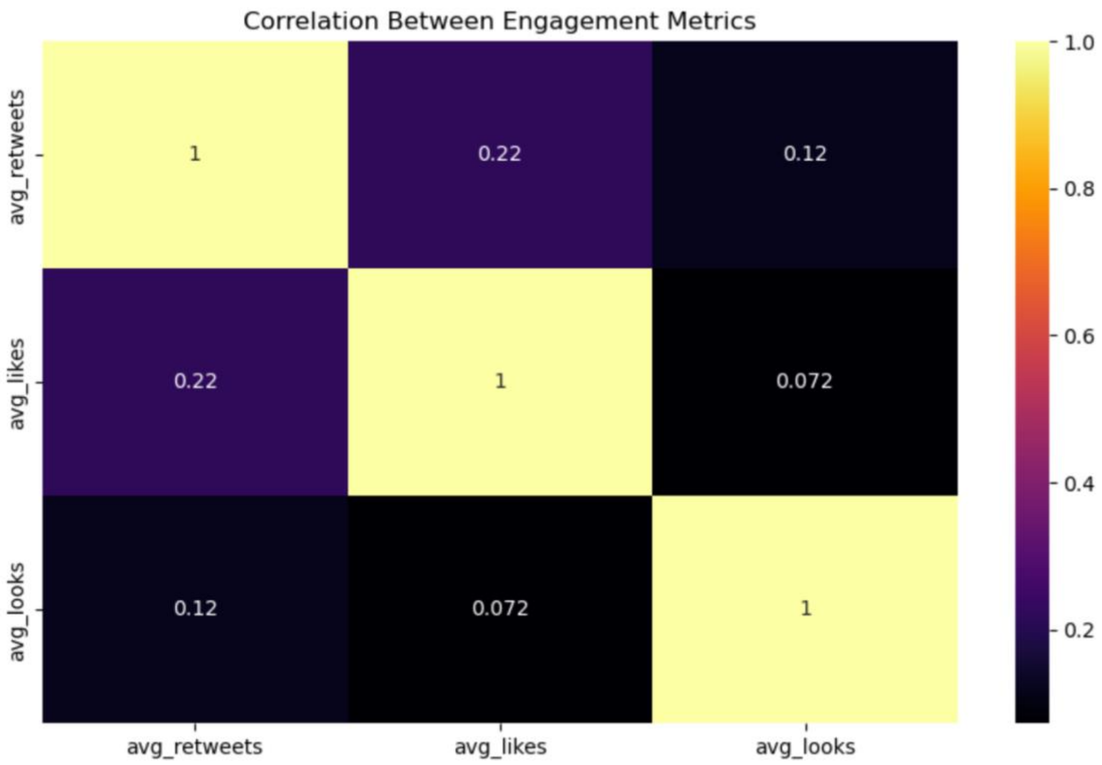


Word Cloud for Negative Sentiment



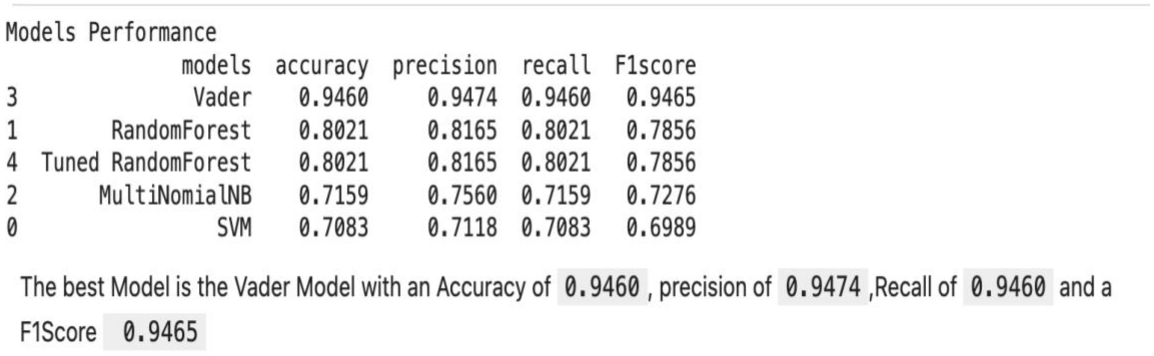

Word Cloud for Neutral Sentiment
letsile tebogo
paris2024 olympic
arshad nadeem
olympics paris2024
gold medal
paris2024 basketball
team usa
world record
nbc peacock
men javelin
network peacock
usabmnt paris2024
parisolympics nbc
usa network
tara daviswoodhall
first olympic
long jump
parisolympics usa
javelin throw
paris2024 olympicgames
bronze medal
noah lyles
olympics2024 paris2024
olympic gold
olympic record
rizki juniansyah
neeraj chopra
united state
silver medalist
olympic medal
gold medalist
men relay
men final
winning silver
winning gold
champ letsile
silver medal
gold paris2024
gold parisolympics
shacarri richardson
olympics olympics2024
sydney mclaughlinlevrone
olympic silver
area record
olympic champ
first time
win gold
nation proud
quincy hall
samsung togetherfortomorrow
track field
medal game
simone bile
made history
peacock parisolympics
osmar olvera
paris olympic
medal olympic
paris2024 omega
final paris2024


Top 20 Bigrams in Olympic Tweets
Bigrams
gold medal
paris2024 olympics
arshad nadeem
letsile tebogo
paris2024 basketball
team usa
olympics paris2024
world record
nbc peacock
win gold
neeraj chopra
men javelin
olympic champion
olympic gold
parisolympics nbc
usabmnt paris2024
olympic record
usa network
network peacock
tara daviswoodhall
0
200
400
600
800
1000
Frequency



MODELING

The modeling process began by preprocessing the data which included cleaning tweets and applying sentiment analysis using VADER. To address class imbalance we employed SMOTE (Synthetic Minority Over-sampling Technique). We then evaluated several machine learning models including Logistic Regression, Support Vector Machine, Random Forest and Naive Bayes using a pipeline approach. Each model was trained and evaluated based on accuracy. The Random Forest model emerged as the best performer with an accuracy of 0.809. Further hyperparameter tuning was performed on this model using GridSearchCV which slightly improved its performance. The final tuned Random Forest model and the TF-IDF vectorizer were saved for future use. Additionally, pre-trained models like VADER and DistilBERT were also evaluated. The VADER model outperformed all others achieving an impressive accuracy of 0.946. Below are the results:



```
Models Performance
                models  accuracy  precision  recall  F1score
3                Vader    0.9460     0.9474  0.9460   0.9465
1         RandomForest    0.8021     0.8165  0.8021   0.7856
4   Tuned RandomForest    0.8021     0.8165  0.8021   0.7856
2        MultiNomialNB    0.7159     0.7560  0.7159   0.7276
0                  SVM    0.7083     0.7118  0.7083   0.6989
```

The best Model is the Vader Model with an Accuracy of `0.9460` , precision of `0.9474` ,Recall of `0.9460` and a F1Score `0.9465`

CONCLUSION

Considering the defined metrics of success, the VADER model exceeded expectations across all categories:

- Accuracy - The VADER model achieved an accuracy of 94.60% surpassing the target range of 85-90%.
- Precision - For the positive class, VADER achieved a precision of 95.96%, and for the negative class, 76.92%. Both of these meet or exceed the target range of 80-90%. The neutral class precision of 96.39% significantly outperformed the 75-85% target. This demonstrates the model's strong ability to correctly identify sentiments particularly for positive and neutral tweets.
- Recall - VADER's recall scores were 95.43% for positive, 84.16% for negative and 95.49% for neutral sentiments. All of these substantially exceed the target range of 75-80% indicating the model's excellent capability in identifying a high proportion of actual sentiments across all classes.
- F1 Score - The overall F1 score of 0.9465 far surpassed the target range of 0.75 to 0.85. This high F1 score reflected a strong balance between precision and recall across all sentiment classes.

In conclusion, the VADER model not only met but significantly exceeded all the defined metrics of success. The success of VADER, a rule-based model specifically designed for social media text, highlights the importance of domain-specific tools in sentiment analysis. While machine learning models like Random Forest showed good performance, they couldn't match the specialized capabilities of VADER in handling the nuances of social media language, particularly in the context of Olympic-related discussions.

RECOMMENDATIONS

1. Implement a real-time sentiment tracking dashboard for organizers and media partners, allowing them to respond quickly to shifts in public opinion.
2. Develop a multi-lingual sentiment analysis capability to cater to the international nature of the Olympics using language-specific versions of VADER where available.
3. Create a sentiment-based alert system for potentially controversial or viral topics enabling rapid response from the communications team.
4. Integrate sentiment analysis results with other data sources (e.g. ticket sales, TV ratings) to provide a comprehensive view of public engagement.
5. Use sentiment trends to guide content creation and social media strategies focusing on themes and athletes that generate positive engagement.
6. Provide regular sentiment reports to sponsors helping them optimize their Olympic-related marketing campaigns.
7. Collaborate with local Paris businesses to use sentiment data for improving visitor experiences during the Olympics.

NEXT STEPS

1. Incorporate Olympics-specific features such as mentions of specific sports, athletes or events to improve classification accuracy.
2. Create a specialized lexicon for VADER that includes Olympic-specific terms and their sentiment associations.
3. Extend the sentiment analysis to multiple social media platforms and news sources for a more comprehensive view.
4. Develop user-friendly and interactive dashboards for stakeholders to explore sentiment data in real-time.
5. Set up a system to compare sentiment trends with previous Olympic events to identify unique characteristics of the Paris Olympics.
6. Develop algorithms to automatically identify and report on significant shifts in sentiment or emerging trends.
7. Offer training sessions for various stakeholders on how to interpret and act upon the sentiment analysis results.
8. Set up infrastructure for continued analysis post-Olympics to track the event's lasting impact on public sentiment towards Paris and the Olympic movement.