












WORLD CUP TWEETS ANALYSIS

BY RAHUL WAGH

TABLE OF CONTENTS

| | |
|---|---------------------------------|
|  | Project Abstract |
|  | Introduction |
|  | Data Preprocessing |
|  | Exploratory Data Analysis (EDA) |
|  | Tweet Text Analysis |
|  | Sentiment Analysis |
|  | Limitations |
|  | Future Work |
|  | Conclusion |

PROJECT ABSTRACT

This project analyzes T20 World Cup 2021 tweets to gain insights into sentiment, user behavior, and key trends. We employ data preprocessing, sentiment analysis, and data visualization to uncover patterns in social media conversations. While presenting our findings, we acknowledge limitations and suggest potential areas for future work. By leveraging this analysis, we aim to enhance our understanding of social media dynamics during major events.

INTRODUCTION

- The project aims to analyze tweets related to the T20 World Cup 2021 and provide insights into user behavior, sentiment, and preferences.
- The dataset used in the project consists of a collection of T20 World Cup 2021 tweets, including information about user profiles, tweet text, hashtags, and more.

DATA PREPROCESSING

- Convert the 'date' column to datetime format.
- Check and display data types of all columns.
- Check for missing values and drop rows with missing values.
- Check for duplicates.
- Display information about the dataset's structure and characteristics.
- Generate summary statistics for both numeric and non-numeric data.

EXPLORATORY DATA ANALYSIS (EDA)

Top Users with Most Tweets

- Create a horizontal bar chart to visualize the top 20 users with the most tweets

Top User Locations

- Create a bar chart to visualize the top 20 user locations

Top User Sources

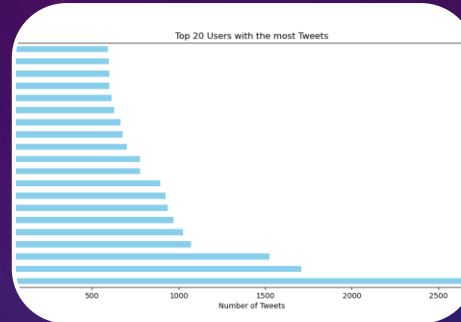
- Create a horizontal bar plot to visualize the top 10 user sources

Top Hashtags

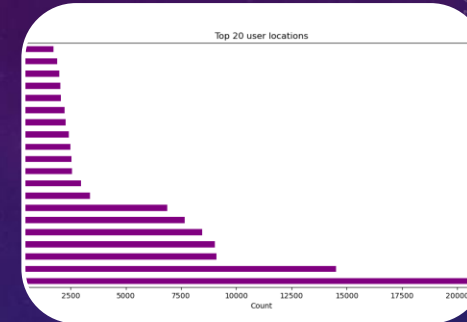
- Create a horizontal bar plot to visualize the top 10 hashtags in the dataset

User Verification Status

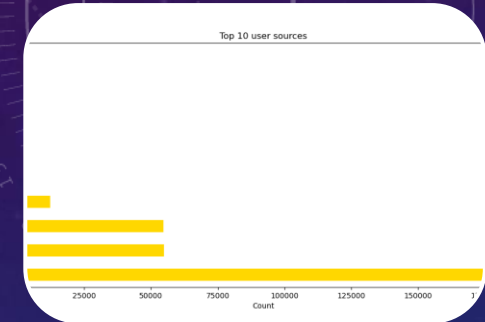
- Display a pie chart to show the distribution of user verification status



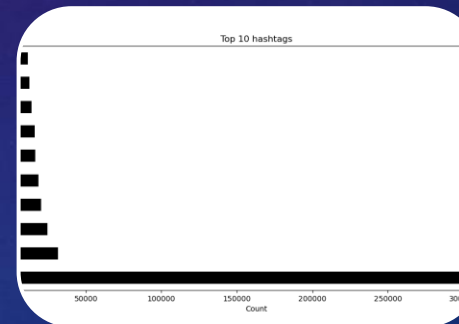
Top Users with Most Tweets



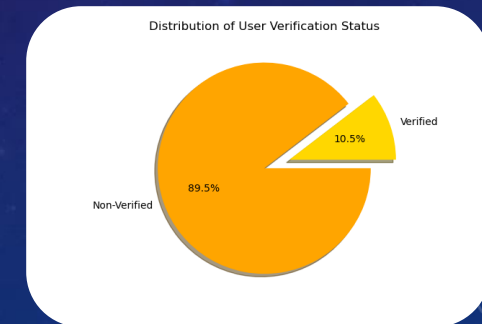
Top User Locations



Top User Sources



Top Hashtags



User Verification Status

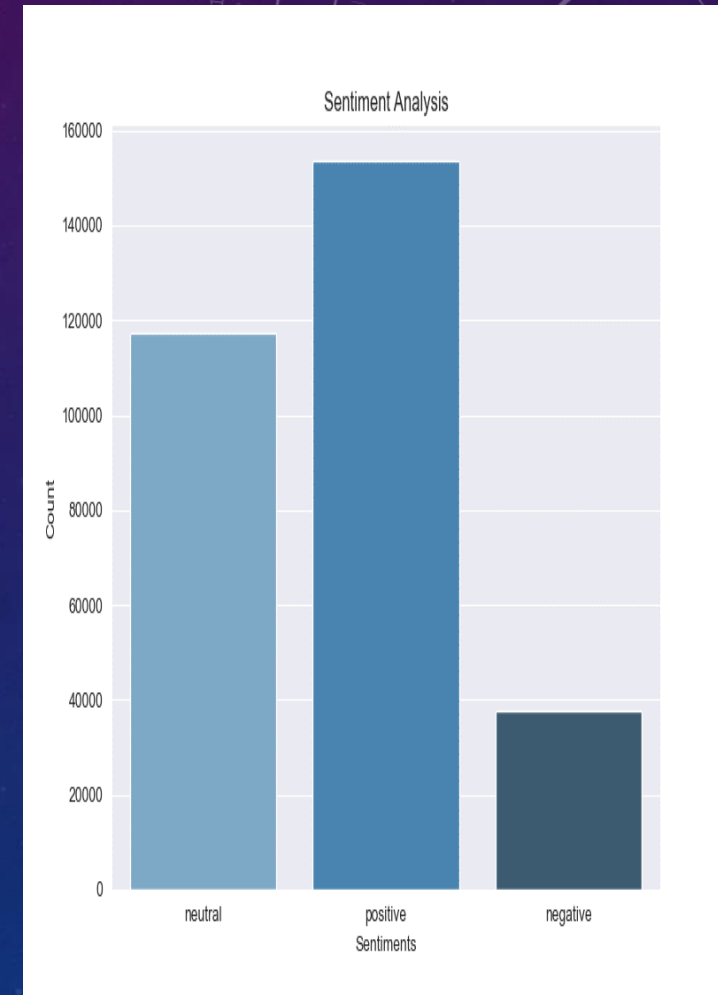
- The text data is preprocessed by removing special characters and links, converting the text to lowercase, and storing the cleaned text in a new column called 'cleaned_text.'
- A word cloud of the tweets is generated to visualize the most common words used in the data.

[illegible][illegible]

- Filter the dataset to select tweets with the user location set as "India"
- Combine cleaned text data into a single string for Indian tweets
- Generate a word cloud to visualize common words in Indian tweets

SENTIMENT ANALYSIS

- Analyze sentiment in tweets and categorize them as 'positive', 'negative', and 'neutral'.
- Create a countplot to visualize the distribution of sentiments.
- Split the dataset into training and testing sets.
- Perform TF-IDF vectorization on the text data.
- Train a Logistic Regression model to predict sentiment.
- Calculate training and testing accuracy.
- Evaluate the model's performance with accuracy, confusion matrix, and classification report.



LIMITATIONS

- **Data Quality:** Social media data can be noisy, with variations in data quality, including spam and non-relevant content.
- **User Biases:** User demographics and participation can lead to biases in the data, potentially affecting the generalizability of our findings.
- **Sentiment Analysis Accuracy:** Sentiment analysis is inherently subjective and may not always accurately capture nuanced sentiments or sarcasm.
- **Machine Learning Model Limitations:** The model's performance depends on the quality and quantity of labeled data. It may require ongoing refinement.
- **Ethical Considerations:** Ensuring ethical data collection and analysis is essential, particularly when dealing with sensitive topics and user privacy.

FUTURE WORK

- **Real-time Analysis:** Implement real-time data collection and analysis for up-to-the-minute insights during events.
- **Advanced Sentiment Analysis:** Explore advanced NLP techniques and sentiment models for more precise sentiment analysis.
- **Model Enhancement:** Fine-tune machine learning models, including deep learning and ensemble methods.
- **Hashtag Tracking:** Develop a dynamic system to track trending hashtags in real-time.

CONCLUSION

- The sentiment analysis successfully categorized tweets into 'positive,' 'negative,' or 'neutral' sentiments.
- The Logistic Regression model demonstrated strong predictive performance, particularly in terms of accuracy, precision, recall, and F1-scores.
- The model's performance was excellent, with high accuracy across all sentiment categories, making it a reliable tool for sentiment analysis in the context of T20 World Cup tweets.
- The project effectively analyzes the T20 World Cup 2021 tweets and provides insights into user behavior, sentiment, and preferences.
- The project can be useful for businesses, marketers, and researchers interested in social media analysis and machine learning.



THANK YOU