TABLE OF CONTENTS

[BUSINESS UNDERSTANDING 2](#_Toc173893716)

[Overview 2](#_Toc173893717)

[Challenges 2](#_Toc173893718)

[Proposed Solutions 3](#_Toc173893719)

[Success Metrics 3](#_Toc173893720)

[Conclusion 4](#_Toc173893721)

[Problem Statement 4](#_Toc173893722)

[Objectives 4](#_Toc173893723)

[**Main Objective 4**](#_Toc173893724)

[**Specific Objectives 4**](#_Toc173893725)

[DATA UNDERSTANDING 5](#_Toc173893726)

[Data Sources 5](#_Toc173893727)

[Datasets 5](#_Toc173893728)

[Relevance of The Data 5](#_Toc173893729)

# 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.

## Success Metrics

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

**85% - 90%**

1. 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.**

1. Recall - The proportion of true positive sentiment predictions out of all actual positives.

**75% - 80% for all sentiment classes.**

1. 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**

1. Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) - Measures how well a model distinguishes between classes.

**> 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

1. APIs - Extract data from social media sites such as Twitter, Facebook and Instagram in the form of posts, tweets, comments and hashtags using their respective APIs.

The focus will be on posts mentioning Paris Olympics, relevant hashtags and location-based data.

1. **Web Scraping -** Extract additional data from comments and discussions from news sites and sports forums such as ESPN and Sports Center

## Datasets

1. **Social media data in the form of tweets, facebook and Instagram posts and comments mentioning the Paris Olympics.**
2. **News articles, comments and replies discussing the various** aspects of the 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. Social media platforms like Twitter, Facebook and Instagram capture immediate reactions, discussions and emotional responses from a global audience thus providing a rich source of unfiltered public sentiment.

The inclusion of location-based data and relevant hashtags allows for more targeted analysis potentially revealing geographical trends and topic-specific sentiments. Complementing this with web scraping of news sites and sports forums like ESPN and Sports Center adds depth to the analysis by incorporating more structured discussions and content.

This combination of data sources offers a comprehensive view of public sentiment ranging from spontaneous reactions on social media to more considered opinions in news comment sections and sports forums.

# DATA PREPARATION

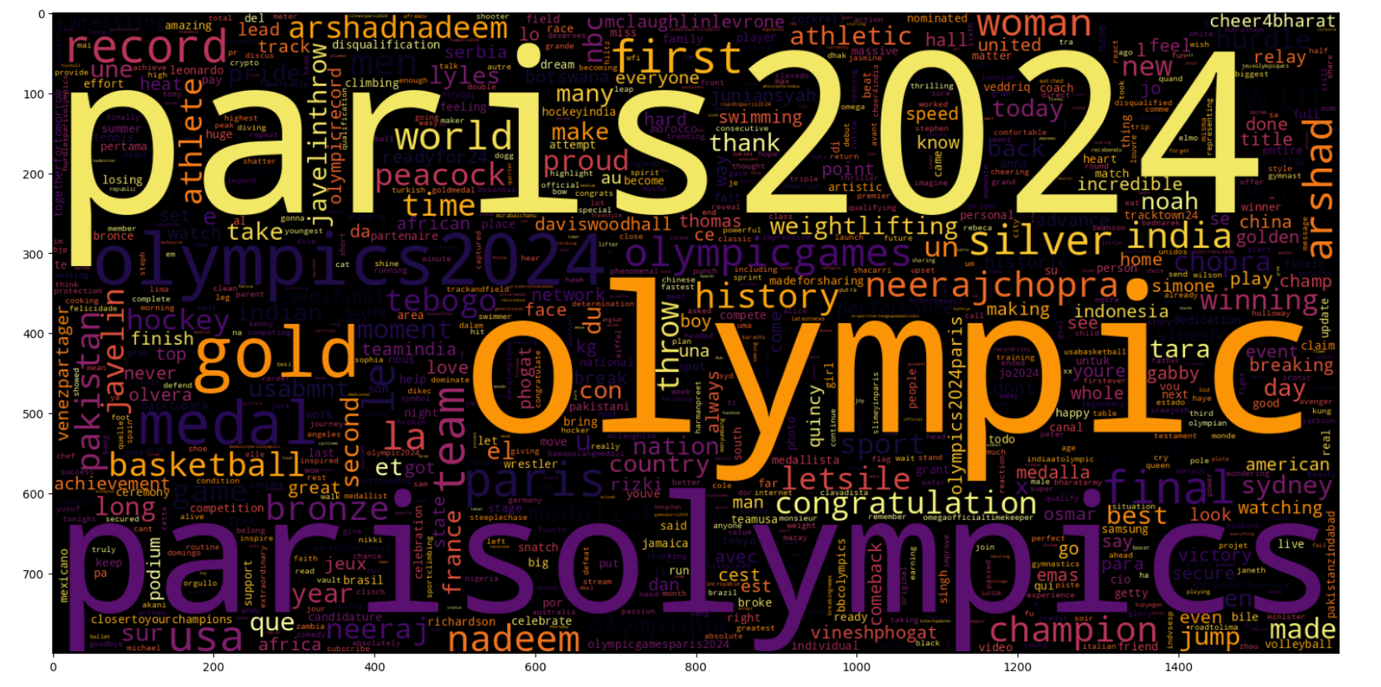
The data processing involved steps to analyze and clean a dataset of tweets related to the 2024 Paris Olympics. The initial dataset was composed of multiple CSV files but was later merged into a single DataFrame. A DataUnderstanding class was created to check the dataset’s shape, columns, unique values,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

Upon examining the dataset we noted that some columns had missing values particularly in the `Tweet\_Timestamp`, `Tweet\_Number\_of\_Reviews`, ``Tweet\_Image\_URL, `Tweet\_Number\_of\_Retweets`, `Tweet\_Number\_of\_Likes` and `Tweet\_Number\_of\_Looks` columns. The dataset also contained a high number of duplicated records which were primarily false positives due to partial 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, mentions, hashtags, punctuation and numbers while retaining relevant tokens. Finally, a word cloud visualization was generated to identify the most frequently used terms in the tweets.



Some tweets appeared to be corrupted with unusual or incorrect words suggesting possible data quality issues that would be resolved for accurate analysis.

## Exploratory Data Analysis