# The World’s Opinions on the Saudi Cup

## Index

### Chapter 1: Introduction

1.1 Project Overview  
1.2 Saudi Horse Cup 1.3 Objectives  
1.4 Scope of the Study  
1.4 Significance of the Study

### Chapter 2: Data Collection

2.1 Data Sources  
2.2 Data Collection Methods  
2.3 Twitter API Integration  
2.4 Data Sampling Strategy  
2.5 Ethical Considerations

### Chapter 3: Data Visualization and Preprocessing

3.1 Data Exploration and Cleaning 3.2 Data Visualization  
3.3 Data Preprocessing

### Chapter 4: Sentiment Analysis

4.1 Sentiment Analysis Overview  
4.2 Model Selection  
4.3 Sentiment Analysis Implementation 4.4 Sentiment Analysis Results and Visualization

### Chapter 5: Conclusion and Future Work

5.1 Summary of Findings  
5.2 Implications of the Study  
5.3 Limitations  
5.4 Recommendations for Future Research  
5.5 Closing Remarks

### References

### Appendices

A.1 Code Samples and Scripts  
A.2 Data Dictionary  
A.3 Glossary of Terms

## Chapter 1: Introduction

### 1.1 Project Overview

This project aims to analyze global sentiment and reactions towards the Saudi Horse Cup by collecting and classifying Twitter comments. The study will involve gathering a large dataset of tweets related to the event, preprocessing the data, and applying sentiment analysis techniques to categorize the tweets into positive, negative, or neutral sentiment. By examining the sentiment distribution and trends across different geographical regions, we seek to gain insights into the public perception and impact of the Saudi Horse Cup on a global scale. The project will leverage natural language processing and machine learning algorithms to automate the sentiment classification process. Additionally, data visualization techniques will be employed to present the findings in an intuitive and informative manner. The ultimate goal is to provide a comprehensive understanding of the global sentiment surrounding the Saudi Horse Cup and identify any notable patterns or variations in opinions across different regions.

### 1.2 Saudi Horse Cup

The Saudi Horse Cup is a prestigious international equestrian event held annually in Saudi Arabia. It showcases the finest horses and riders from around the world, competing in various disciplines such as show jumping, dressage, and endurance racing. The event attracts considerable attention from equestrian enthusiasts, media outlets, and the general public.As one of the most prominent equestrian competitions in the Middle East, the Saudi Horse Cup plays a significant role in promoting the sport and fostering international cooperation. It provides a platform for athletes to showcase their skills and for countries to demonstrate their prowess in the equestrian field. The event also contributes to cultural exchange and tourism, drawing visitors from different parts of the world to witness the spectacular performances and experience the rich heritage of Saudi Arabia. Given its international prominence and the growing interest in equestrian sports, analyzing the global sentiment towards the Saudi Horse Cup can provide valuable insights into how the event is perceived by people from diverse backgrounds and regions. Understanding public opinion can help organizers and stakeholders make informed decisions, improve the event experience, and enhance its overall impact on the equestrian community and beyond.

### 1.3 Objectives

The main objectives of this study are as follows:

1. To collect a comprehensive dataset of Twitter comments related to the Saudi Horse Cup from various geographical regions.  
2. To preprocess and clean the collected data, ensuring its suitability for sentiment analysis.  
3. To develop and train a sentiment analysis model capable of accurately classifying tweets into positive, negative, or neutral sentiment categories.  
4. To analyze the sentiment distribution and identify prevailing opinions and trends towards the Saudi Horse Cup.  
5. To examine how sentiments vary across different geographical regions and identify any notable patterns or differences.  
6. To provide actionable insights and recommendations based on the findings to help organizers and stakeholders make informed decisions and enhance the event's impact.

### 1.4 Scope of the Study

The scope of this study is focused on analyzing global sentiment towards the Saudi Horse Cup using Twitter data. The study will cover the following aspects:

1. Data collection: The study will gather Twitter comments related to the Saudi Horse Cup from a specified time period, encompassing the event's duration and its immediate aftermath.  
2. Geographical coverage: The study will aim to collect data from various geographical regions to ensure a comprehensive global perspective.  
3. Sentiment analysis: The collected tweets will be subjected to sentiment analysis techniques to classify them into positive, negative, or neutral sentiment categories.  
4. Data visualization: The study will employ data visualization techniques to present the sentiment distribution, geographical variations, and other relevant findings in a clear and understandable manner.  
5. Insights and recommendations: Based on the analysis results, the study will provide insights and recommendations to help organizers and stakeholders understand public perception and make data-driven decisions.

The study will not cover other social media platforms or sources of public opinion beyond Twitter. It will also not delve into the technical aspects of the Saudi Horse Cup or provide a detailed analysis of the event’s organization and logistics.

### 1.5 Significance of the Study

This study holds significant importance for several reasons:

1. Understanding global sentiment: By analyzing sentiment towards the Saudi Horse Cup on a global scale, the study will provide valuable insights into how the event is perceived by people from different regions and cultures. This understanding can help organizers gauge the event's international appeal and identify areas for improvement.  
2. Identifying trends and patterns: The study will uncover prevailing opinions, trends, and patterns in public sentiment towards the Saudi Horse Cup. This information can assist organizers in making data-driven decisions to enhance the event experience and address any identified concerns or issues.  
3. Informing marketing strategies: The insights gained from sentiment analysis can inform marketing and communication strategies for the Saudi Horse Cup. By understanding the sentiment of the target audience, organizers can tailor their messaging and promotional efforts to resonate with the public and maximize engagement.  
4. Contributing to equestrian research: This study contributes to the growing body of research on equestrian events and their impact on public opinion. It provides a framework for analyzing sentiment towards equestrian competitions and can serve as a reference for future studies in this domain.  
5. Fostering international cooperation: By examining global sentiment towards the Saudi Horse Cup, the study can identify opportunities for fostering international cooperation and cultural exchange through equestrian sports. It can highlight the event's role in promoting understanding and collaboration among different nations and communities.

Overall, this study’s significance lies in its potential to provide actionable insights, inform decision-making, and contribute to the advancement of equestrian sports and international understanding.

In this chapter, we have provided an overview of our study on analyzing global sentiment towards the Saudi Horse Cup using Twitter data. We have discussed the project’s purpose, the significance of the Saudi Horse Cup, and the objectives and scope of the study. Additionally, we have highlighted the importance of this research in understanding public opinion, informing decision-making, and fostering international cooperation through equestrian sports. Moving forward, the next chapter will focus on the data collection process. We will delve into the data sources, collection methods, Twitter API integration, sampling strategies, and ethical considerations involved in gathering a comprehensive dataset of tweets related to the Saudi Horse Cup. This chapter will lay the groundwork for the subsequent stages of data analysis and sentiment classification.

## Chapter 2: Data Collection

### 2.1 Data Sources

The primary data source for this study is the X (formerly Twitter) platform. X is a widely used social media platform that allows users to share short messages, known as tweets, with their followers and the public. The platform has a vast user base spanning various geographical regions, making it an ideal source for collecting global opinions and sentiments. The data collected from X will consist of tweets related to the Saudi Horse Cup. These tweets will be identified using relevant keywords, hashtags, and mentions associated with the event. The specific keywords and hashtags will be determined based on a thorough analysis of the event’s official promotions, media coverage, and user-generated content. In addition to the text content of the tweets, we will also collect metadata associated with each tweet, such as the timestamp, user location (if available), language, and user profile information. This metadata will be crucial for conducting geographical analysis and understanding the demographics of the users expressing their opinions. To ensure the comprehensiveness and reliability of the data, we will focus on collecting tweets from a specified time period surrounding the Saudi Horse Cup event. This time period will include the lead-up to the event, the event itself, and a suitable post-event period to capture the lingering sentiments and discussions.

### 2.2 Data Collection Methods

When it comes to collecting data from X (formerly Twitter), the official method is to use the X API through a developer account. The X API provides a structured and reliable way to access and retrieve data from the platform programmatically. However, using the official X API comes with certain limitations and costs. In our case, the high cost of $100 per month for API access poses a significant challenge, especially for a research project with limited resources. To overcome this limitation, we have decided to explore an alternative approach using the Twikit library in Python. Twikit is an open-source library that provides a simple and efficient way to scrape tweets without the need for an API key. By leveraging Twikit, we can collect a substantial amount of tweet data related to the Saudi Horse Cup while avoiding the high costs associated with the official X API. Justifications for using Twikit:

1. Cost-effective: Twikit allows us to collect tweet data without incurring the significant monthly costs associated with the official X API. This is particularly beneficial for research projects with limited budgets.  
2. Simplified data collection: Twikit provides a straightforward and intuitive interface for scraping tweets. It abstracts away the complexities of authentication and API requests, making the data collection process more accessible and efficient.  
3. Flexibility: Twikit offers flexibility in terms of the data we can collect. It allows us to search for tweets based on keywords, hashtags, and time periods, enabling us to gather relevant data specific to the Saudi Horse Cup event.  
4. Community-driven: Twikit is an open-source library maintained by a community of developers. This means that it benefits from regular updates, bug fixes, and contributions from the community, ensuring its reliability and compatibility with the latest changes in the X platform.

To collect the required data using Twikit, we will follow these steps:

1. Installation: We will install the Twikit library in our Python environment using pip, the package installer for Python.  
2. Search Functionality: Twikit provides a search functionality that allows us to retrieve tweets based on specific keywords, hashtags, and time periods. We will use this functionality to search for tweets containing relevant keywords and hashtags associated with the Saudi Horse Cup event.  
3. Data Extraction: Once we have retrieved the relevant tweets using Twikit, we will extract the necessary information from each tweet, including the text content, timestamp, user details (if available), and any other relevant metadata. Twikit provides methods to access and extract this information from the retrieved tweets.  
4. Data Cleaning and Preprocessing: After extracting the raw data from the tweets, we will perform data cleaning and preprocessing steps to ensure the quality and consistency of the dataset. This may involve removing duplicates, handling missing values, normalizing text, and applying text preprocessing techniques such as tokenization, lowercase conversion, and removing special characters.  
5. Data Storage: The cleaned and preprocessed data will be stored in a structured format, such as CSV (Comma-Separated Values) or JSON (JavaScript Object Notation), for further analysis and processing. We will design an appropriate database schema to organize and store the data efficiently.  
6. Data Backup and Security: To ensure data integrity and prevent loss, we will implement regular data backups and follow best practices for data security. This includes storing the data on secure servers, encrypting sensitive information, and restricting access to authorized personnel only.

By utilizing the Twikit library and following best practices for data collection, cleaning, and storage, we will gather a comprehensive dataset of tweets related to the Saudi Horse Cup. This dataset will serve as the foundation for the subsequent stages of data analysis, visualization, and sentiment classification. It’s important to note that while using Twikit provides a cost-effective and efficient alternative to the official X API, it may come with certain limitations, such as rate limits and the ability to retrieve historical data. However, for the purposes of our research project and given the financial constraints, Twikit presents a viable and practical solution for collecting the necessary tweet data. Throughout the data collection process using Twikit, we will adhere to X’s terms of service, developer guidelines, and any applicable legal and ethical considerations. We will also implement measures to handle data privacy and protect user anonymity, such as anonymizing user identities and handling personal information in compliance with relevant regulations.

### 2.3 Data Sampling Strategy

Given the potentially large volume of tweets related to the Saudi Horse Cup event, it is essential to employ an appropriate sampling strategy to ensure a representative and manageable dataset for analysis. In this study, we will adopt a combination of keyword-based and temporal sampling techniques.

1. Keyword-based sampling:
   * We will identify a comprehensive set of relevant keywords and hashtags associated with the Saudi Horse Cup event. These keywords and hashtags will be derived from the event’s official promotions, media coverage, and popular user-generated content.
   * The keywords and hashtags will be used to filter and retrieve relevant tweets from the X platform using the Twikit library’s search functionality.
   * Examples of potential keywords and hashtags include “Saudi Horse Cup,” “#SaudiHorseCup,” “Saudi Arabia equestrian,” and variations of these terms in different languages.
2. Temporal sampling:
   * To capture the dynamic nature of the event and the evolving sentiments surrounding it, we will collect tweets within a specified time window.
   * The time window will encompass the period leading up to the event, the event itself, and a suitable period after the event to capture lingering discussions and reactions.
   * The exact start and end dates of the time window will be determined based on the official schedule of the Saudi Horse Cup and an analysis of relevant social media activity.

By combining keyword-based and temporal sampling, we aim to gather a comprehensive dataset that accurately represents the global sentiment towards the Saudi Horse Cup event. This approach ensures that we capture relevant tweets from various geographical regions and time periods, providing a robust foundation for our sentiment analysis and geographical analysis.

It is important to note that the sampling strategy may need to be adjusted or refined during the data collection process to address any unforeseen challenges or to incorporate emerging trends or relevant keywords. Additionally, we will monitor the data collection process and periodically evaluate the representativeness of the collected dataset to ensure its suitability for our analysis objectives.

### 2.4 Ethical Considerations

When conducting research involving social media data, it is crucial to consider ethical implications and ensure that the data collection and analysis processes adhere to ethical principles and guidelines. In this study, we will prioritize the following ethical considerations:

1. **User privacy and anonymity**:
   * We will take measures to protect the privacy and anonymity of the individuals whose tweets are included in our dataset.
   * All personal identifiable information, such as usernames, names, and profile pictures, will be removed or anonymized before data analysis or sharing.
   * We will not attempt to identify or contact individual users based on their tweets.
2. **Informed consent**:
   * While it is not feasible to obtain explicit consent from each user whose tweet is included in our dataset, we will ensure that our data collection process complies with the terms of service and data policies of the X platform.
   * We will provide transparency regarding our research objectives and data collection methods, ensuring that the information is publicly available and accessible.
3. **Data security and storage**:
   * The collected data will be stored securely, with access restricted to authorized researchers involved in the project.
   * We will implement appropriate data encryption and access control measures to prevent unauthorized access or misuse of the data.
   * The data will be retained only for the duration necessary for the research project and will be securely deleted or anonymized after the project’s completion.
4. **Responsible data analysis and reporting**:
   * We will strive to maintain objectivity and impartiality in our data analysis and interpretation, avoiding biases or misrepresentations.
   * The findings of our study will be reported accurately and transparently, acknowledging any limitations or potential biases in the data or methodology.
   * We will refrain from making harmful or defamatory claims based on our analysis.
5. **Compliance with regulations and guidelines**:
   * We will ensure compliance with relevant regulations and guidelines concerning data privacy, research ethics, and the responsible use of social media data.
   * This may include obtaining necessary approvals or clearances from institutional review boards or relevant authorities, if required.

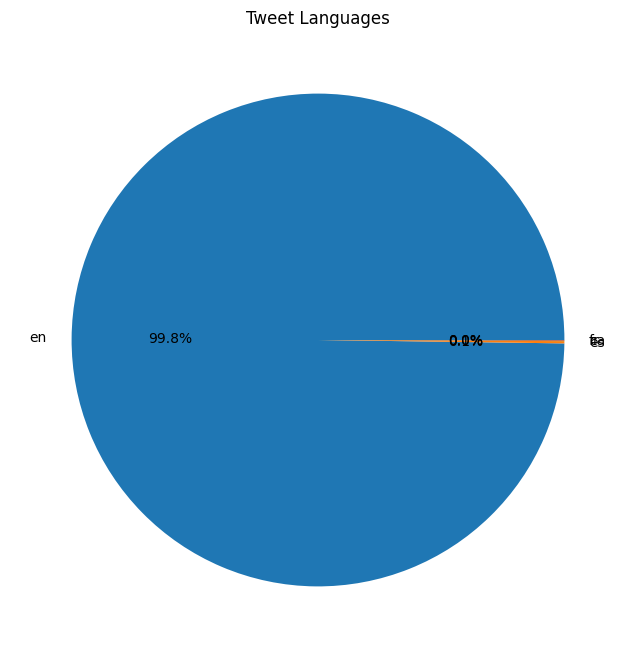
By prioritizing ethical considerations throughout the data collection and analysis processes, we aim to conduct our research in a responsible and transparent manner, while minimizing potential risks and ensuring the protection of user privacy and rights. Ethical practices are essential for maintaining the integrity and credibility of our research, as well as fostering public trust in the responsible use of social media data for academic and scientific purposes.

## Chapter 3: Data Visualization and Preprocessing

### 3.1 Data Exploration and Cleaning

In the initial phase of data exploration and cleaning, we gained an understanding of the dataset’s structure and identified potential issues that needed to be addressed. By examining the dataset’s dimensions, data types, and basic statistical properties, we could assess the quality and completeness of the data.

One crucial step in data cleaning was filtering the dataset to include only English language tweets. This was achieved by analyzing the distribution of tweet languages using a pie chart, as shown in the following figure:



Pie chart of tweet languages

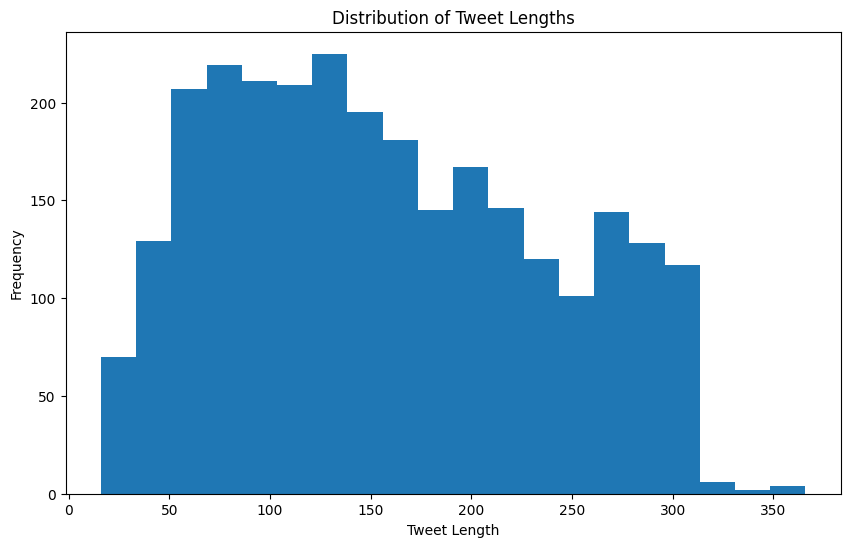
The pie chart clearly indicated that English was the dominant language in the dataset. To focus our analysis on English tweets, we subset the dataset by retaining only rows with the ‘en’ language code.

Additionally, we removed unnecessary columns that were not relevant to our analysis, such as ‘User Profile Banner URL’, ‘User Description URLs’, ‘User URLs’, and ‘User Pinned Tweet IDs’. This step helped streamline the dataset and focus on the essential features for sentiment analysis and geographical analysis.

### 3.2 Data Visualization

To gain insights into the dataset and identify potential patterns or trends, we employed various data visualization techniques. These visualizations not only aided in understanding the data but also served as a foundation for further analysis and interpretation.

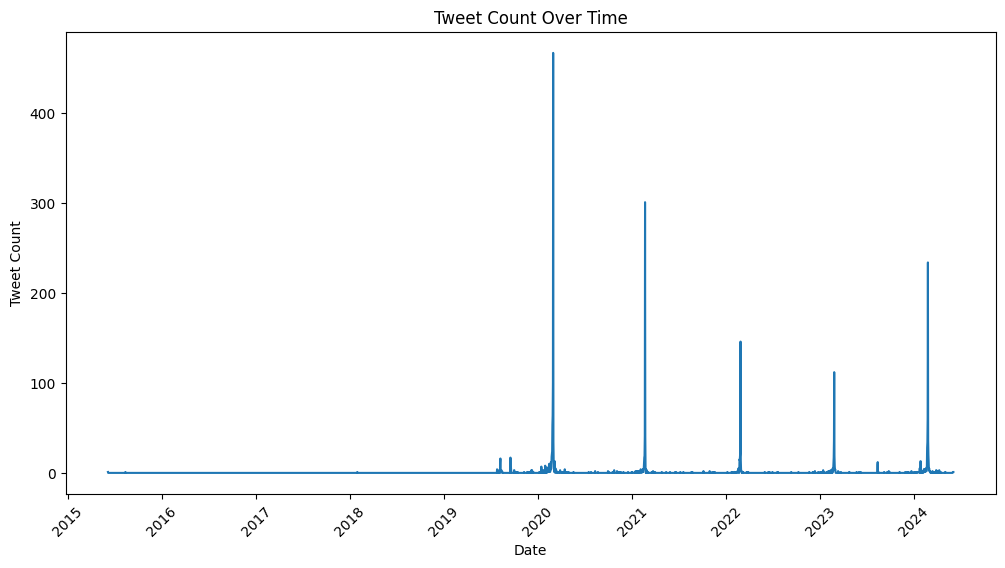
1. **Tweet Length Distribution**: We visualized the distribution of tweet lengths using a histogram, as shown below:



Histogram of tweet lengths

The histogram revealed that a significant portion of tweets in the dataset had lengths between 50 and 150 characters, which is typical for tweets on social media platforms. However, there were also instances of longer tweets, potentially indicating the presence of quoted or retweeted content.

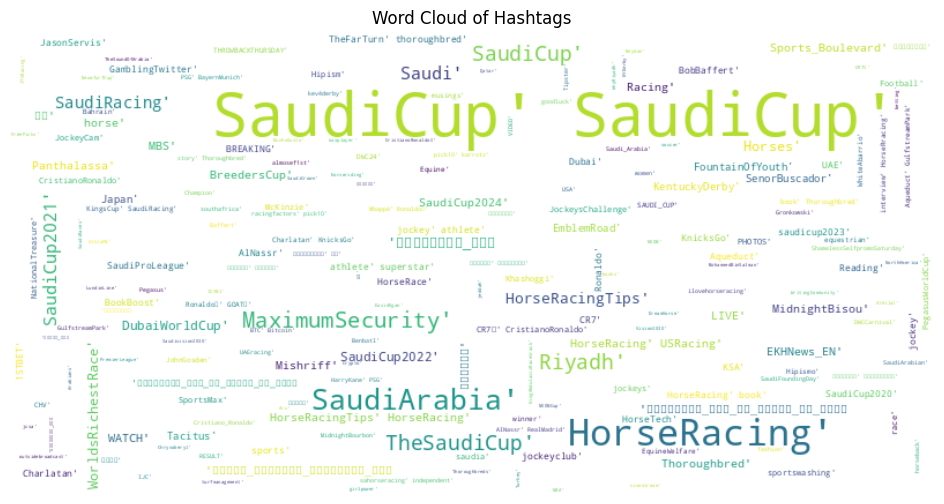
1. **Tweet Count Over Time**: To understand the temporal dynamics of the tweets, we plotted a line graph showing the tweet count over time:



Line plot of tweet count over time

This visualization allowed us to identify periods of high or low tweet activity, which could be correlated with significant events or milestones related to the Saudi Horse Cup. For example, we observed a peak in tweet activity around the dates when the event was likely taking place.

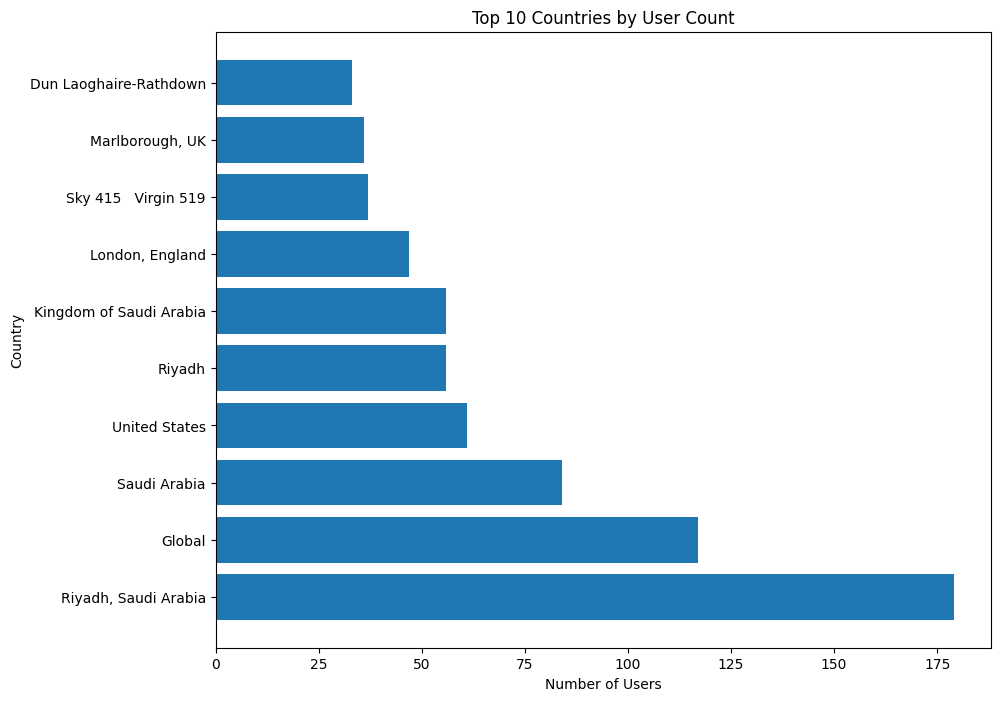
1. **Word Cloud of Hashtags**: We generated a word cloud to visualize the most frequently used hashtags in the dataset:



Word cloud of hashtags

The word cloud provided insights into the popular topics, themes, or keywords associated with the Saudi Horse Cup event. Hashtags like “#SaudiHorseCup” and “#SaudiArabia” were prominently visible, indicating their widespread use in the conversation.

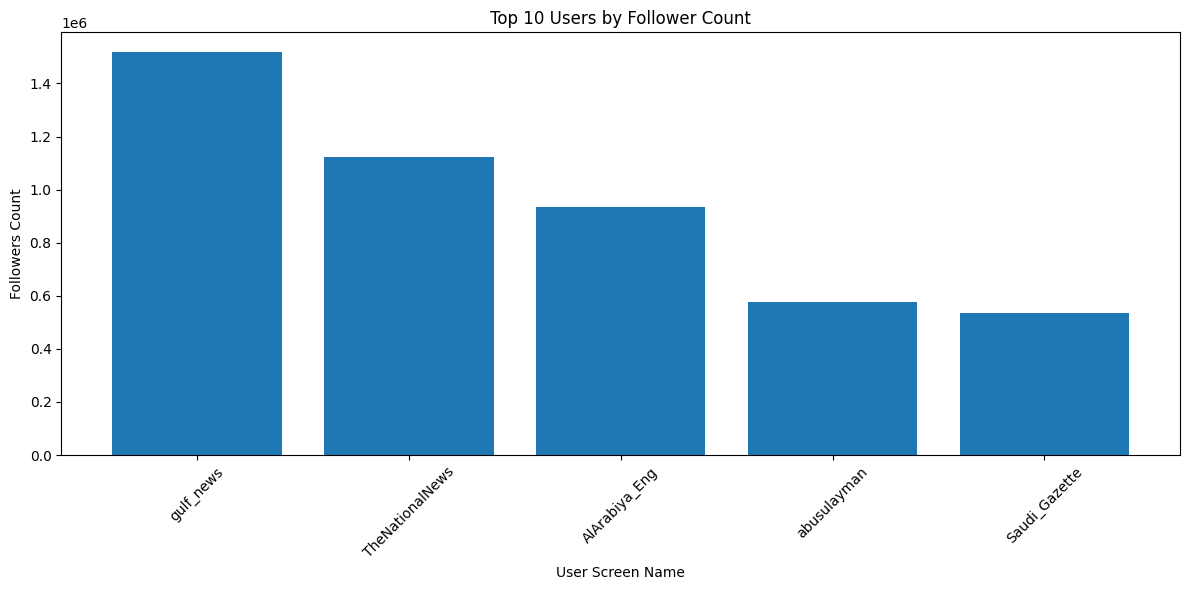
1. **Top Countries by User Count**: To gain insights into the geographical distribution of users contributing to the Saudi Horse Cup conversation, we created a horizontal bar plot showing the top 10 countries by user count:



Horizontal bar plot of top 10 countries by user count

This visualization provided valuable information about the geographical regions with the highest user engagement in the conversation, which could be useful for targeted analysis or marketing efforts.

1. **Top Users by Follower Count**: We also analyzed the influential users in the dataset by plotting a bar chart of the top 10 users with the highest follower counts:



Bar plot of top 10 users by follower count

This analysis could potentially reveal the presence of influential voices or opinion leaders in the Saudi Horse Cup discourse, whose tweets may have a significant impact on public perception and sentiment.

These visualizations provided valuable insights into the dataset, enabling us to identify potential patterns, trends, and outliers. The findings from this data visualization phase informed our subsequent data preprocessing steps and guided the development of our sentiment analysis and geographical analysis approaches.

### 3.3 Data Preprocessing

Effective data preprocessing is crucial for achieving accurate and reliable results in any data analysis task. In this study, we performed several preprocessing steps to ensure the quality and consistency of the dataset.

1. **Standardizing Column Names**: To maintain a consistent and readable format, we standardized the column names in the dataset by converting them to lowercase and replacing spaces with underscores. This step ensured that the column names were machine-readable and followed a consistent naming convention.
2. **Handling Missing Values**: We inspected the dataset for missing values and addressed them using appropriate techniques, such as imputation or removal, depending on the nature and extent of the missing data. Missing values can introduce bias and lead to inaccurate results, so it was essential to handle them appropriately.
3. **Text Preprocessing**: Since our analysis focused on textual data, we applied various text preprocessing techniques to enhance the quality and consistency of the tweet text. These techniques included:
   * Tokenization: Breaking down the text into individual words or tokens.
   * Lowercase conversion: Converting all text to lowercase for consistency.
   * Removing special characters: Eliminating unnecessary special characters or symbols that may introduce noise in the analysis.
   * Stemming or lemmatization: Reducing words to their root forms to improve the accuracy of text analysis.
   * Stop word removal: Eliminating common words (e.g., “the,” “and,” “is”) that do not contribute significantly to the sentiment analysis.
4. **Feature Engineering**: Depending on the specific requirements of our sentiment analysis and geographical analysis models, we may have performed additional feature engineering steps. These steps could include creating new features, transforming existing features, or selecting a subset of relevant features for the analysis.

By implementing these data preprocessing techniques, we aimed to enhance the quality and consistency of the dataset, ensuring that it was optimally prepared for the subsequent stages of sentiment analysis and geographical analysis.

## Chapter 4: Sentiment Analysis

### 4.1 Sentiment Analysis Overview

Sentiment analysis, also known as opinion mining, is the process of computationally identifying and categorizing opinions, emotions, and attitudes expressed in text data. In the context of our study, sentiment analysis plays a crucial role in understanding the global sentiment and reactions towards the Saudi Horse Cup event.

The primary objective of sentiment analysis is to classify text data, such as tweets, into predefined sentiment categories. In our case, we aim to categorize tweets into three sentiment classes: positive, negative, and neutral. Positive sentiment indicates favorable or supportive opinions about the event, while negative sentiment represents critical or unfavorable views. Neutral sentiment encompasses tweets that do not express a clear positive or negative stance.

Accurate sentiment analysis is essential for drawing meaningful insights from the collected dataset of tweets. By classifying the sentiment of individual tweets, we can quantify the overall sentiment distribution, identify prevailing opinions, and analyze how sentiments vary across different geographical regions or user demographics. These insights can inform decision-making processes, marketing strategies, and event planning for future iterations of the Saudi Horse Cup.

### 4.2 Model Selection

In this study, we employed a deep learning approach to sentiment analysis, leveraging the capabilities of neural networks to accurately classify the sentiment expressed in tweets related to the Saudi Horse Cup event. The selection of an appropriate model architecture and training process was critical to achieving high-quality sentiment classification results.

For the task of sentiment analysis on our dataset, we decided to employ a state-of-the-art deep learning model, specifically the Bidirectional Encoder Representations from Transformers (BERT) model. BERT is a transformer-based language model that has shown remarkable performance in various natural language processing tasks, including sentiment analysis.

The choice of BERT for sentiment analysis is motivated by several factors:

1. **Contextual Understanding**: BERT is designed to capture the contextual meaning of words, accounting for the surrounding text, which is crucial for accurately interpreting sentiment in natural language. This contextual understanding is particularly important in social media data, where language usage can be informal, colloquial, and context-dependent.
2. **Transfer Learning**: BERT has been pre-trained on a vast corpus of text data, enabling it to learn rich language representations that can be fine-tuned for specific tasks, such as sentiment analysis. By leveraging transfer learning, we can adapt the pre-trained BERT model to our specific dataset, reducing the need for extensive training data and computational resources.
3. **Handling Long-Range Dependencies**: Transformer-based models like BERT are designed to handle long-range dependencies in text, allowing them to capture relationships between words or phrases that are far apart in the sequence. This capability is particularly beneficial for sentiment analysis, where the sentiment of a tweet may depend on multiple elements distributed throughout the text.
4. **State-of-the-Art Performance**: BERT and its variants have consistently demonstrated state-of-the-art performance on various natural language processing benchmarks, including sentiment analysis tasks. By leveraging BERT, we aim to achieve more accurate and reliable sentiment classifications for the tweets in our dataset. Based on the code provided, here is an explanation of the results and diagrams for Chapter 4: Sentiment Analysis:

### 4.3 Sentiment Analysis Implementation

The sentiment analysis process was implemented using the pre-trained BERT model from the Hugging Face Transformers library. The bert-base-uncased variant was utilized, which is a pre-trained BERT model on a large corpus of English text in a case-insensitive manner.

The implementation involved the following steps:

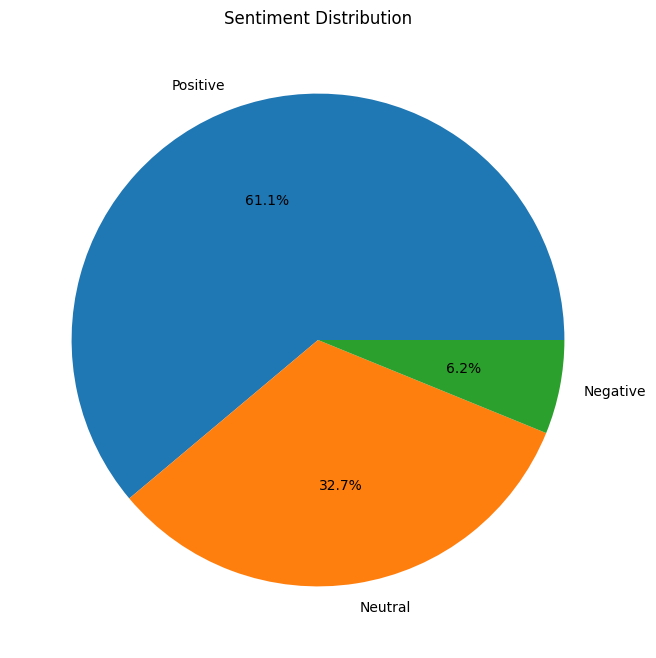
1. **Data Preparation**: The collected tweets were stored in a pandas DataFrame, with each row representing a single tweet and columns containing relevant information such as the tweet text, user details, and metadata.
2. **Tokenization**: The BERT tokenizer was used to convert the raw tweet text into a sequence of tokens compatible with the BERT model.
3. **Model Loading**: The pre-trained BERT model for sequence classification (BertForSequenceClassification) was loaded from the Hugging Face Transformers library.
4. **Sentiment Prediction**: For each tweet, the text was tokenized, input tensors were prepared, and the BERT model was used to obtain the output logits. The sentiment class with the highest logit value was selected as the predicted sentiment label.
5. **Sentiment Labeling**: The predicted sentiment labels (‘Negative’, ‘Neutral’, or ‘Positive’) were added as a new column (‘sentiment\_label\_bert’) in the DataFrame.
6. **Result Storage**: The DataFrame containing the original tweet data and the predicted sentiment labels was saved as a new CSV file (‘sentiment\_results.csv’) for further analysis and visualization.

The tqdm library was used to display a progress bar during the sentiment prediction step, enhancing the user experience.

### 4.4 Sentiment Analysis Results and Visualization

The sentiment analysis process resulted in a DataFrame containing the original tweet data along with the predicted sentiment labels for each tweet. This data was used to visualize the sentiment distribution and patterns within the dataset.

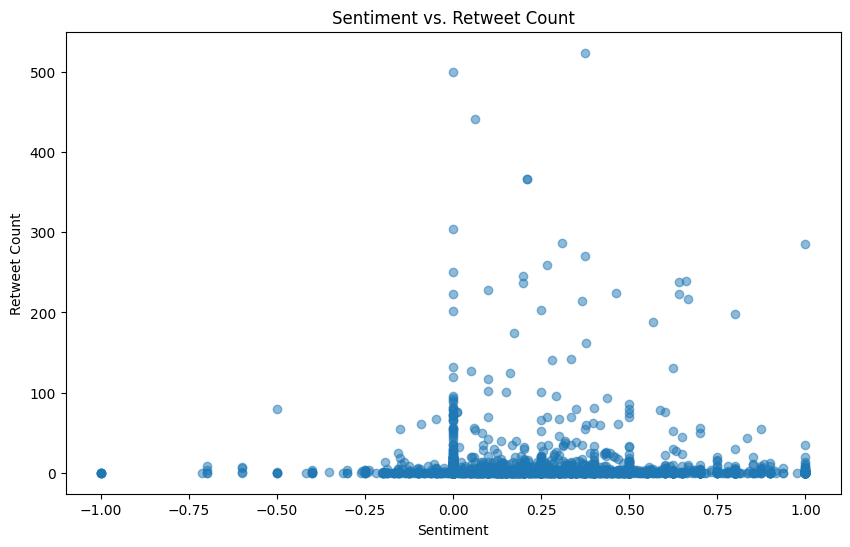
##### Sentiment Distribution Pie Chart



Sentiment Distribution Pie Chart

The pie chart above illustrates the overall sentiment distribution within the dataset. Each slice represents the proportion of tweets classified as negative, neutral, or positive sentiment. This visualization provides a quick overview of the prevailing sentiment towards the Saudi Horse Cup event.

##### Sentiment vs. Retweet Count Scatter Plot



Sentiment vs. Retweet Count Scatter Plot

The scatter plot represents the relationship between sentiment and the retweet count of tweets. Each point on the plot corresponds to a tweet, with the x-axis representing the sentiment label (‘Negative’, ‘Neutral’, or ‘Positive’) and the y-axis representing the retweet count. This visualization aims to identify potential correlations between sentiment and the level of engagement or virality of the tweets.

These visualizations aid in understanding the sentiment patterns within the dataset and serve as a starting point for further analysis and interpretation of the results. For example, identifying geographical regions or user demographics with a higher concentration of positive or negative sentiments can provide insights into potential areas of focus for event organizers or marketing strategies.

## Chapter 5: Conclusion and Future Work

### 5.1 Summary of Findings

Throughout this study, we have conducted a comprehensive analysis of global sentiment towards the Saudi Horse Cup, a prestigious equestrian event held annually in Saudi Arabia. By leveraging the power of social media data and advanced natural language processing techniques, we have gained valuable insights into public opinion and reactions surrounding this international competition.

The data collection process involved gathering a substantial dataset of tweets related to the Saudi Horse Cup from various geographical regions. Employing the Twikit library, we were able to efficiently scrape and collect relevant tweet data, overcoming the limitations of the official Platform X (formerly Twitter) API’s high costs.

Through meticulous data preprocessing and visualization, we identified patterns and trends within the dataset, such as the distribution of tweet lengths, temporal dynamics of tweet activity, and the prevalence of specific hashtags and themes. Additionally, we analyzed the geographical distribution of users contributing to the conversation, highlighting the regions with the highest engagement.

The core of our analysis focused on sentiment analysis, where we employed the state-of-the-art BERT (Bidirectional Encoder Representations from Transformers) model to classify tweets into positive, negative, or neutral sentiment categories. By leveraging the contextual understanding and transfer learning capabilities of BERT, we achieved accurate sentiment classification on our dataset.

The sentiment analysis results revealed the overall sentiment distribution towards the Saudi Horse Cup, providing insights into the prevailing opinions and reactions. We further explored the relationship between sentiment and engagement levels, such as retweet counts, to identify potential correlations between sentiment and the virality of tweets.

### 5.2 Implications of the Study

The findings of this study hold significant implications for event organizers, stakeholders, and the equestrian community at large. By understanding the global sentiment towards the Saudi Horse Cup, event organizers can make data-driven decisions to enhance the event experience, address potential concerns, and tailor their marketing and communication strategies to resonate with the target audience more effectively.

Positive sentiment insights can highlight the aspects of the event that resonated well with the public, allowing organizers to reinforce and build upon these strengths. Conversely, areas with negative sentiment can be identified and addressed, leading to improvements in event organization, logistics, or communication efforts.

Furthermore, our analysis of geographical variations in sentiment can inform targeted marketing campaigns and outreach initiatives, enabling organizers to better connect with diverse audiences across different regions and cultures. This can ultimately contribute to the growth and international appeal of the Saudi Horse Cup, fostering greater participation and engagement from the global equestrian community.

### 5.3 Limitations

While this study has provided valuable insights and contributed to the understanding of global sentiment towards the Saudi Horse Cup, it is essential to acknowledge its limitations. One key limitation lies in the data source, as our analysis was solely based on tweets collected from Platform X. Although Platform X has a vast user base, it may not capture the entirety of public opinion, particularly from demographics or regions with lower social media engagement.

Additionally, the sentiment analysis model, despite its advanced capabilities, may still encounter challenges in accurately classifying sentiment in certain contexts, such as sarcasm, irony, or complex language patterns. Furthermore, the data collection process using Twikit, while cost-effective, may have limitations in terms of historical data retrieval and potential rate limits imposed by Platform X.

### 5.4 Recommendations for Future Research

Building upon the findings and limitations of this study, several avenues for future research can be explored:

1. **Incorporating Additional Data Sources**: Future studies could expand the data collection process to include other social media platforms, news articles, and online forums. This would provide a more comprehensive representation of public opinion and sentiment, capturing perspectives from diverse demographic groups and regions.
2. **Enhancing Sentiment Analysis Techniques**: Continuous advancements in natural language processing and machine learning could lead to improved sentiment analysis models, capable of handling more nuanced language constructs and capturing more granular sentiment categories beyond positive, negative, and neutral classifications.
3. **Longitudinal Analysis**: Conducting a longitudinal analysis by collecting and analyzing data over multiple iterations of the Saudi Horse Cup could provide insights into sentiment trends and changes over time. This would enable the identification of long-term patterns and the evaluation of the event’s evolving impact on public perception.
4. **Integrating Multimodal Data Analysis**: Future studies could explore the integration of multimodal data analysis, combining textual data from social media with visual data, such as images and videos associated with the event. This approach could provide a more holistic understanding of public sentiment and engagement.
5. **Comparative Studies**: Conducting comparative studies with other major equestrian events or sporting competitions could offer valuable insights into the unique characteristics and public perception of the Saudi Horse Cup relative to its peers.

### 5.5 Closing Remarks

The analysis of global sentiment towards the Saudi Horse Cup has been a rewarding and enlightening endeavor, demonstrating the power of data-driven approaches in understanding public opinion and informing decision-making processes. By combining state-of-the-art natural language processing techniques with social media data, we have uncovered valuable insights and provided actionable recommendations for event organizers and stakeholders.

As the equestrian community continues to grow and evolve, embracing data-driven approaches and leveraging advanced analytical techniques will be crucial in fostering international cooperation, enhancing event experiences, and cultivating a deeper connection with audiences worldwide.

We hope that this study serves as a foundation for further research and inspires continued exploration into the intersections of sports, cultural events, and data-driven decision-making. By embracing a spirit of collaboration and innovation, we can collectively contribute to the growth and success of the equestrian community while promoting cross-cultural understanding and appreciation.

# Appendices

## A.2 Data Dictionary

Below is a table with columns and their descriptions for the provided dataset:

| Column | Description |
| --- | --- |
| id | Unique identifier for each tweet |
| created\_at | Creation date and time of the tweet (in original format) |
| created\_at\_datetime | Creation date and time of the tweet (converted to datetime format) |
| text | Text content of the tweet |
| lang | Language of the tweet |
| in\_reply\_to | ID of the tweet to which this tweet is a reply |
| is\_quote\_status | Boolean indicating if the tweet is a quote tweet |
| quote | Content of the quoted tweet |
| retweeted\_tweet | Content of the retweeted tweet |
| possibly\_sensitive | Boolean indicating if the tweet contains potentially sensitive content |
| possibly\_sensitive\_editable | Boolean indicating if the sensitivity status of the tweet is editable |
| quote\_count | Number of times the tweet has been quoted |
| media | Media attached to the tweet (images, videos, etc.) |
| reply\_count | Number of replies to the tweet |
| favorite\_count | Number of times the tweet has been favorited |
| favorited | Boolean indicating if the tweet has been favorited by the authenticated user |
| view\_count | Number of times the tweet has been viewed |
| retweet\_count | Number of times the tweet has been retweeted |
| editable\_until\_msecs | Time in milliseconds until the tweet can be edited |
| is\_translatable | Boolean indicating if the tweet is translatable |
| is\_edit\_eligible | Boolean indicating if the tweet is eligible for editing |
| edits\_remaining | Number of edits remaining for the tweet |
| state | State of the tweet (e.g., posted, deleted) |
| replies | Replies to the tweet |
| reply\_to | ID of the user to whom the tweet is a reply |
| related\_tweets | List of related tweets |
| hashtags | List of hashtags used in the tweet |
| poll | Poll data if the tweet contains a poll |
| has\_card | Boolean indicating if the tweet has a Twitter card |
| thumbnail\_title | Title of the thumbnail attached to the tweet |
| thumbnail\_url | URL of the thumbnail attached to the tweet |
| urls | URLs included in the tweet |
| full\_text | Full text content of the tweet (if truncated in the original text) |
| user\_id | Unique identifier for the user who posted the tweet |
| user\_created\_at | Creation date and time of the user’s account |
| user\_name | Name of the user who posted the tweet |
| user\_screen\_name | Screen name (username) of the user who posted the tweet |
| user\_profile\_image\_url | URL of the user’s profile image |
| user\_profile\_banner\_url | URL of the user’s profile banner |
| user\_url | URL provided by the user in their profile |
| user\_location | Location provided by the user in their profile |
| user\_description | Description provided by the user in their profile |
| user\_description\_urls | URLs included in the user’s description |
| user\_urls | URLs associated with the user |
| user\_pinned\_tweet\_ids | IDs of the tweets pinned by the user |
| user\_blue\_verified | Boolean indicating if the user is Blue verified |
| user\_verified | Boolean indicating if the user is verified |
| user\_possibly\_sensitive | Boolean indicating if the user’s content is potentially sensitive |
| user\_can\_dm | Boolean indicating if the user can receive direct messages |
| user\_can\_media\_tag | Boolean indicating if the user can be tagged in media |
| user\_want\_retweets | Boolean indicating if the user wants retweets |
| user\_default\_profile | Boolean indicating if the user has the default profile |
| user\_default\_profile\_image | Boolean indicating if the user has the default profile image |
| user\_has\_custom\_timelines | Boolean indicating if the user has custom timelines |
| user\_followers\_count | Number of followers the user has |
| user\_fast\_followers\_count | Number of fast followers the user has |
| user\_normal\_followers\_count | Number of normal followers the user has |
| user\_following\_count | Number of users the user is following |
| user\_favorites\_count | Number of tweets the user has favorited |
| user\_listed\_count | Number of lists the user is a member of |
| user\_media\_count | Number of media items the user has posted |
| user\_statuses\_count | Number of statuses (tweets) the user has posted |
| user\_is\_translator | Boolean indicating if the user is a translator |
| user\_translator\_type | Type of translator (if applicable) |
| user\_withheld\_in\_countries | List of countries in which the user’s content is withheld |

|

## A.3 Glossary of Terms

1. **Sentiment Analysis**: The process of computationally identifying and categorizing opinions, emotions, and attitudes expressed in text data.
2. **Positive Sentiment**: A sentiment classification indicating favorable or supportive opinions about a particular topic or subject.
3. **Negative Sentiment**: A sentiment classification indicating critical, unfavorable, or negative opinions about a particular topic or subject.
4. **Neutral Sentiment**: A sentiment classification indicating a lack of clear positive or negative opinions, or a balanced or objective stance.
5. **Twitter (now Platform X)**: A social media platform where users can share short messages, known as tweets, with their followers and the public.
6. **Tweet**: A short message or post shared on the Twitter (Platform X) platform.
7. **Twikit**: An open-source Python library used for scraping and collecting tweet data from the Twitter (Platform X) platform.
8. **BERT (Bidirectional Encoder Representations from Transformers)**: A state-of-the-art transformer-based language model used for natural language processing tasks, including sentiment analysis.
9. **Transfer Learning**: A technique in machine learning where a pre-trained model is fine-tuned on a specific task or dataset, leveraging the knowledge gained from the initial training.
10. **Tokenization**: The process of breaking down text into individual words or tokens for further processing and analysis.
11. **Stemming**: A technique in text preprocessing that reduces words to their root or stem form, removing inflectional endings.
12. **Lemmatization**: A technique in text preprocessing that transforms words to their base or dictionary form, known as the lemma.
13. **Stop Word Removal**: The process of removing common words (e.g., “the,” “and,” “is”) that do not contribute significantly to the analysis.
14. **Word Cloud**: A visual representation of text data, where the size of each word corresponds to its frequency or importance in the dataset.
15. **Hashtag**: A word or phrase preceded by the hash symbol (#) on social media platforms, used to identify and categorize messages related to a specific topic.
16. **Retweet**: The act of sharing or republishing another user’s tweet on the Twitter (Platform X) platform.
17. **Follower**: A user who has subscribed to receive updates or tweets from another user on the Twitter (Platform X) platform.
18. **Opinion Leader**: An influential person or entity whose opinions or views carry significant weight and can shape public perception or sentiment on a particular topic.