عنوان المشروع : " آراء العالم حول كأس السعودية للخيل "

الغرض من هذا المشروع هو تحليل المشاعر من خلال ايجاد صورة واضحة عن تأثير هذه الفعالية على العالم ومعرفة ردود الفعل تجاه الفعالية كتصنيفها إلى سلبة أو إيجابية او محايدة وفهم التوزيع الجغرافي للآراء.

**ملاحظات:**

جمع بيانات من تغريدات على توتير في الفترة ما بين

لازم أن تكون التغريدات باللغة الانجليزية

وأن لا تتحاوز 5000 صف

كيف يمكننا الاستفادة من هذه البيانات من خلال تدريب الذكاء الاصطناعي عليها مثل تصنيفها إلى سلبة أو إيجابية او محايدة يمكننا الاستعانة بنموذج BERT

وما هو الغرض من المشروع والفوائد وما هو نطاق المشروع؟

**خطوات المشروع:**

* جمع البيانات
* إجراء تحليل وتصور عن البيانات
* عمل نموذج بيرت
* كتابة تقرير وورد بشباتر
* كتابة ورقة علمية مختصرة
* عمل عرض بوربوينت

Write me an index for the documentation for this project. It must contain at least five chapters. Each chapter contains a group of sections such as (1.1, 1.2, 1.3, …etc.)

Write in

* 1. Project Overview
  2. Saudi Horse Cup

## Title : “The World’s Opinions on the Saudi Horse Cup”

Purpose

The purpose of this project is to analyze global sentiment and reactions towards the Saudi Horse Cup by collecting and classifying Twitter comments. This will help in understanding the event’s impact, identifying prevailing opinions, and examining how these sentiments vary geographically. By training an AI model on this data, we aim to categorize the sentiments into negative, positive, or neutral, providing insights into public perception.

## "The World's Opinions on the Saudi Horse Cup"

### Chapter 1: Introduction

* 1. Project Overview
  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 Annotation and Visualization

3.1 Data Annotation

3.2 Sentiment Distribution Charts

3.3 Word Clouds and Keyword Analysis

3.4 Time-Series Analysis of Tweets

### Chapter 4: Data Preprocessing

4.1 Data Cleaning

4.2 Handling Missing Values

4.3 Data Normalization

4.4 Text Preprocessing Techniques

### Chapter 5: Sentiment Analysis

5.1 Sentiment Analysis Overview

5.2 Model Selection

5.3 Training the Sentiment Analysis Model

5.4 Model Evaluation

5.5 Sentiment Categorization

### Chapter 6: Geographical Analysis

6.1 Geolocation Data Extraction

6.2 Mapping Tweets to Geographic Regions

6.3 Regional Sentiment Trends

6.4 Comparative Analysis of Geographic Sentiments

6.5 Challenges in Geographical Analysis

### Chapter 7: Conclusion and Future Work

7.1 Summary of Findings

7.2 Implications of the Study

7.3 Limitations

7.4 Recommendations for Future Research

7.5 Closing Remarks

### References

### Appendices

A.1 Code Samples and Scripts

A.2 Data Dictionary

A.3 Glossary of Terms

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.

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.

2.4 Data Sampling Strategy

2.4 Ethical Considerations