Applied Artificial Intelligence

Project Title: "***China: Web Scraping and Sentiment Analysis with Streamlit App"***

Submitted By:

Pavithra Sevakula [70572200011]

Submitted To:

Prof. Rajesh Prabhakar

Department of Computer Science Engineering – Data Science

SVKM'S NMIMS HYDERABAD Deemed to-be UNIVERSITY

1. Executive Summary

This report presents a comprehensive sentiment analysis of content from the Wikipedia page on China, utilizing natural language processing (NLP) techniques and machine learning models. The project demonstrates how text data can be scraped, processed, analyzed, and visualized to extract meaningful insights about sentiment patterns and trends in informational content about China.

The analysis revealed a distribution of 265 positive, 231 neutral, and 114 negative sentences in the Wikipedia content, indicating an overall balanced representation with a slightly positive tilt. Through machine learning classification, the project achieved a 76% accuracy rate in predicting sentiment, with Logistic Regression emerging as the best-performing model with an F1 score of 0.7352.

The interactive Streamlit application developed as part of this project allows users to input any text about China and receive real-time sentiment analysis results, comparing predictions from our trained machine learning model with baseline TextBlob analysis.

---

2. Introduction

Sentiment analysis is a rapidly growing field within Natural Language Processing (NLP) that aims to identify and extract subjective information from text. In an era of information abundance, understanding the emotional tone and sentiment of content about specific topics is crucial for researchers, policymakers, and the general public.

This project focuses on analyzing the sentiment of text related to China, using Wikipedia as a comprehensive and relatively neutral information source. China, as one of the world's most influential countries, is discussed extensively across various domains, making it an ideal subject for sentiment analysis. By examining how China is portrayed in educational contexts like Wikipedia, we can gain insights into the prevailing narratives and tones used in objective, encyclopedic sources.

The project combines web scraping, text preprocessing, sentiment analysis using TextBlob, and advanced machine learning techniques to build a robust sentiment classification system. The final product is an interactive web application that allows users to analyze the sentiment of any text related to China, providing both machine learning-based predictions and TextBlob analysis for comparison.

3. Project Objectives

The main objectives of this project are:

1. To collect and analyze textual data about China from Wikipedia
2. To develop a robust text preprocessing pipeline for sentiment analysis
3. To explore the distribution of sentiments in Wikipedia's content about China
4. To build and evaluate multiple machine learning models for sentiment classification
5. To compare traditional lexicon-based sentiment analysis (TextBlob) with machine learning approaches
6. To develop an interactive web application that allows users to analyze the sentiment of any China-related text
7. To provide insights into the portrayal of China in educational, encyclopedic contexts

---

4. Methodology

The project followed a systematic approach to analyze sentiment in the Wikipedia content about China:

Data Collection

- Web scraping the Wikipedia page on China using the BeautifulSoup library
- Extracting paragraph text from the main content area

Data Preprocessing

- Cleaning text by removing references, special characters, and extra whitespace
- Tokenizing text into sentences using NLTK
- Filtering out short segments (less than 10 characters)

Sentiment Analysis

- Using TextBlob to calculate polarity scores for each sentence
- Classifying sentences as positive, negative, or neutral based on polarity values

- Creating a structured dataframe of sentences with their sentiment classifications

Text Feature Analysis

- Tokenizing text into words and removing stopwords
- Generating word frequency counts and visualizing using WordCloud
- Identifying the most common words in the content

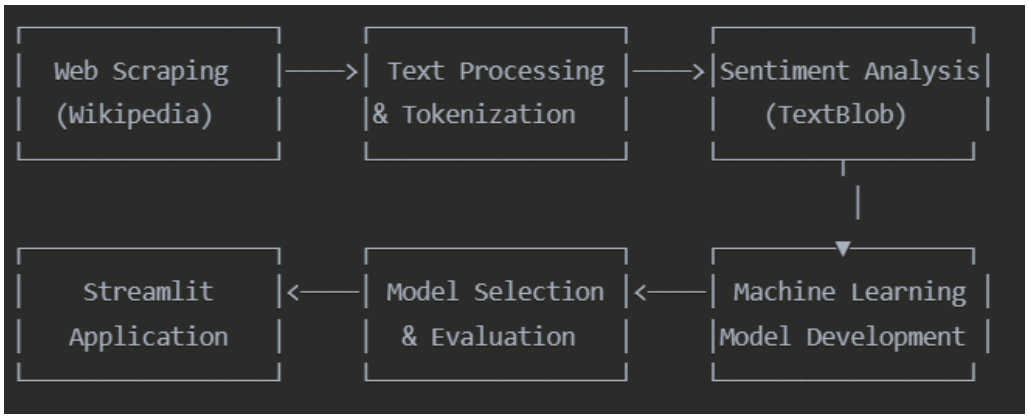
Machine Learning Model Development

- Converting text to numerical features using TF-IDF vectorization
- Removing neutral sentiments to focus on binary classification
- Applying SMOTE technique to balance class distribution
- Splitting data into training and testing sets (70/30 split)
- Training and evaluating multiple classification models:
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - Gradient Boosting
    - Naive Bayes
    - K-Nearest Neighbors

Interactive Application Development

- Creating a Streamlit web application for real-time sentiment analysis
- Implementing user interface with input text area and analysis results
- Displaying visual representations of sentiment predictions
- Providing comparison between machine learning model and TextBlob results

Workflow Diagram



---

5. Results and Analysis

Text Analysis

The text analysis of the Wikipedia content on China yielded the following insights:

- **Sentence Distribution**: Out of 610 total sentences:
  - Positive: 265 sentences (43.4%)
  - Neutral: 231 sentences (37.9%)
  - Negative: 114 sentences (18.7%)
- **Word Frequency**: The most frequent words included:
  - "china" (324 occurrences)
  - "chinese" (134 occurrences)
  - "world" (112 occurrences)
  - "largest" (72 occurrences)
  - "country" (64 occurrences)

Word Cloud Visualization

Figure 1: Word cloud visualization of the most frequent terms in the China Wikipedia page

6. Model Performance

Multiple machine learning models were trained and evaluated, with the following results:

Model	Accuracy	F1 Score
Logistic Regression	0.7632	0.7352
Decision Tree	0.6842	0.6899
Random Forest	0.7632	0.7077
Gradient Boosting	0.7368	0.7209
Naive Bayes	0.6842	0.6684
K-Nearest Neighbors	0.6053	0.5979

Logistic Regression emerged as the best-performing model, with an accuracy of 76.32% and an F1 score of 0.7352. This model was selected for deployment in the interactive Streamlit application.

Performance Metrics Visualization

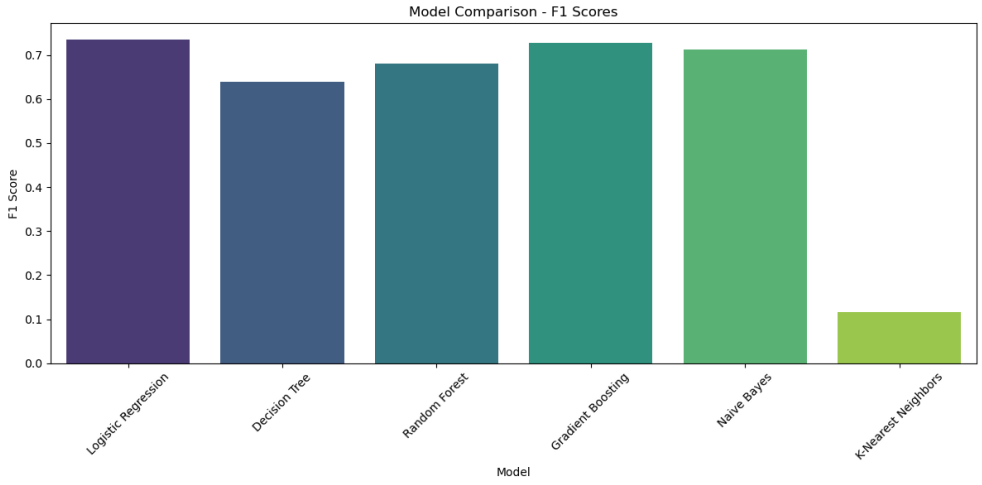


Figure 2: Comparison of accuracy and F1 scores across different machine learning models

7. Streamlit Application

The Streamlit application provides an interactive interface for users to analyze the sentiment of any text related to China. The application features:

1. A text input area for entering sentences about China
2. Machine learning model prediction with visualization of probability scores
3. Comparison with TextBlob sentiment analysis
4. Downloadable visualization of sentiment probability
5. Information about the model and data source

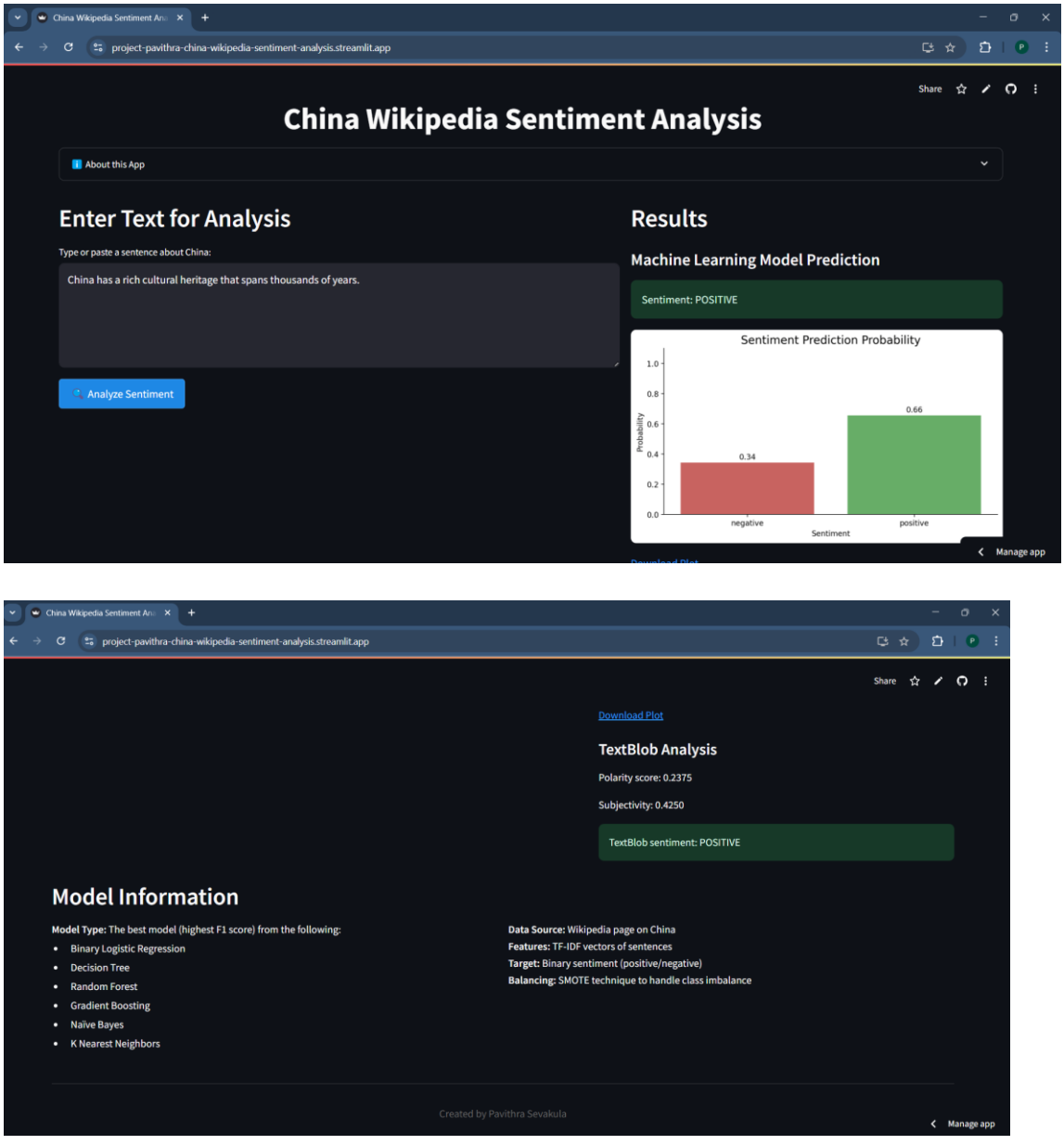
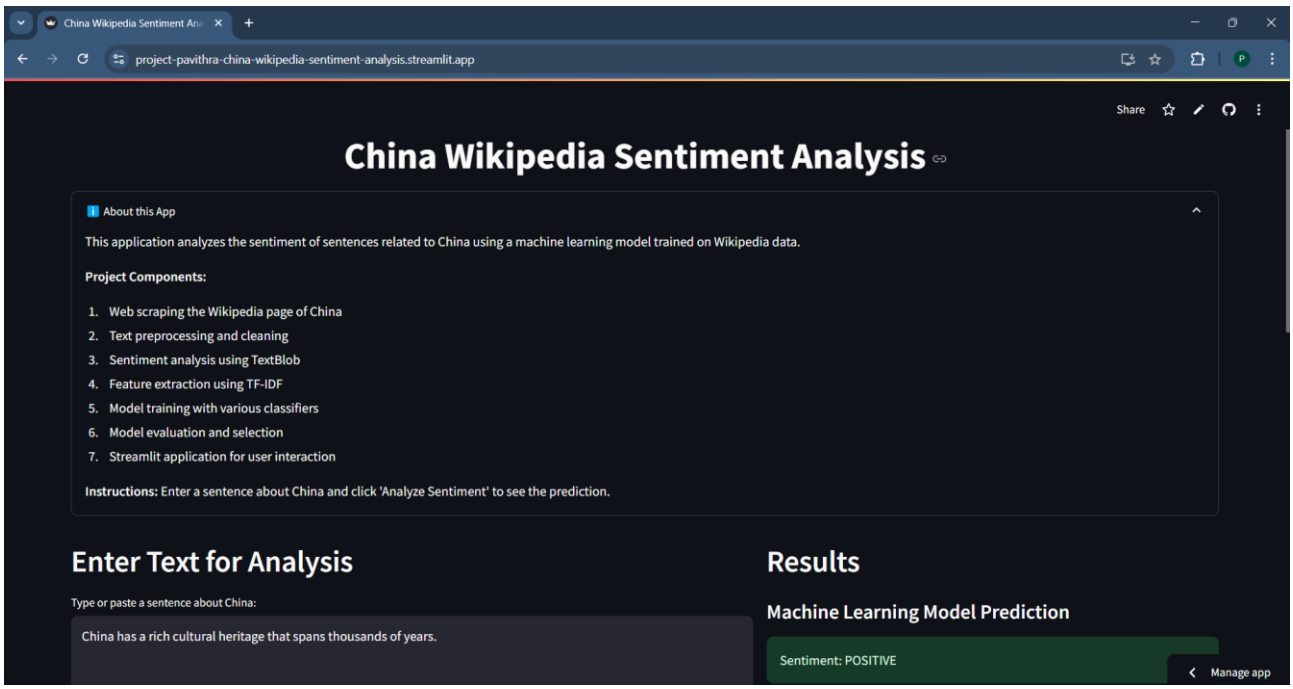


Figure 3: Screenshot of the Streamlit application interface showing sentiment analysis results



Sentiment Distribution

The Wikipedia content about China shows a predominantly positive or neutral tone, with only 18.7% of sentences classified as negative. This aligns with Wikipedia's goal of providing objective information, though the slight positive tilt (43.4% positive sentences) may indicate some inherent bias in the content.

8. Model Performance Analysis

The Logistic Regression model outperformed other classification algorithms, suggesting that linear relationships between words and sentiment are more effective for this dataset than complex tree-based or ensemble methods. The relatively small dataset size (379 sentences after removing neutral content) may have affected the performance of more complex models like Random Forest and Gradient Boosting, which typically require larger training sets.

---

9. Challenges and Limitations

Several challenges were encountered during the project:

1. **Class Imbalance**: The original dataset had significantly fewer negative sentences compared to positive and neutral ones, requiring the application of SMOTE for balancing.
2. **Contextual Understanding**: Machine learning models struggle with understanding context, sarcasm, and implicit meanings, which can affect sentiment classification accuracy.
3. **Subjectivity in Ground Truth**: Using TextBlob's polarity scores as the initial ground truth introduces some subjectivity, as TextBlob itself is not perfectly accurate.
4. **Domain-Specific Language**: The Wikipedia page contains specialized terms and references to historical events that may not be well-captured by general-purpose sentiment analysis tools.

---

10. Conclusion and Future Work

This project successfully implemented a sentiment analysis pipeline for analyzing Wikipedia content about China, achieving a 76.32% accuracy in sentiment classification. The interactive Streamlit application provides a user-friendly interface for real-time sentiment analysis of text related to China.

Future work could include:

1. **Expanded Data Sources**: Incorporating multiple Wikipedia pages or different sources about China for a more comprehensive analysis.
2. **Advanced NLP Techniques**: Implementing more sophisticated methods like BERT or other transformer-based models for improved context understanding.

3. **Temporal Analysis**: Analyzing how sentiment in Wikipedia content about China has changed over time by examining historical page versions.
4. **Multilingual Analysis**: Extending the analysis to Wikipedia pages about China in different languages to compare sentiment across cultural perspectives.
5. **Topic-Specific Sentiment**: Breaking down sentiment analysis by topics (economy, history, politics, culture) for more granular insights.

---

11. Technical Implementation

The project was implemented using Python with the following libraries:

- pandas, numpy for data manipulation
- BeautifulSoup, requests for web scraping
- NLTK for text processing
- TextBlob for baseline sentiment analysis
- scikit-learn for machine learning models
- imbalanced-learn for handling class imbalance
- Streamlit for web application development
- matplotlib, seaborn, wordcloud for visualization

---

12. Links

- GitHub Repository: https://github.com/Pavithrasevakula/china-wikipedia-sentiment-analysis.git
- Streamlit Application: https://project-pavithra-china-wikipedia-sentiment-analysis.streamlit.app/