Applied Artificial Intelligence

Netherlands: Web Scraping and Sentiment Analysis with Streamlit App

Submitted By:

Malde Saicharan: [70572200033]

Submitted To: Prof. Rajesh Prabhakar

SVKM’S NMIMS HYDERABAD

Project Overview

This project implements a comprehensive sentiment analysis system focused on content from the Netherlands Wikipedia page. The project showcases a complete data science pipeline including web scraping, text preprocessing, exploratory data analysis, model training, and deployment via a Streamlit web application. The work demonstrates proficiency in natural language processing, machine learning, and web application development.

Technical Implementation Analysis

Technologies Used Programming Languages & Libraries

The project leverages various tools and libraries for different stages of processing, model training, and deployment:

• Python: Core programming language used for all processing

• Streamlit: Web framework for building an interactive app

• Scikit-learn: Machine learning library for training models

• Pandas & NumPy: Data manipulation and numerical operations

• NLTK (Natural Language Toolkit): Text preprocessing and tokenization

• Pickle: For saving and loading trained models

Data Collection and Preprocessing

The project begins with web scraping of the Netherlands Wikipedia page using Beautiful Soup, extracting paragraphs from the main content. The preprocessing workflow is thorough:

Netherlands - Wikipedia

- Citation removal (e.g., [1], [2])
- Special character and digit removal
- White space normalization
- Sentence tokenization using NLTK
- Word tokenization and stopword removal

This rigorous preprocessing creates a clean dataset suitable for sentiment analysis and machine learning applications.

Exploratory Data Analysis

The EDA process includes:

1. **Sentiment Distribution Analysis**: Using TextBlob to categorize sentences as positive, negative, or neutral based on polarity scores

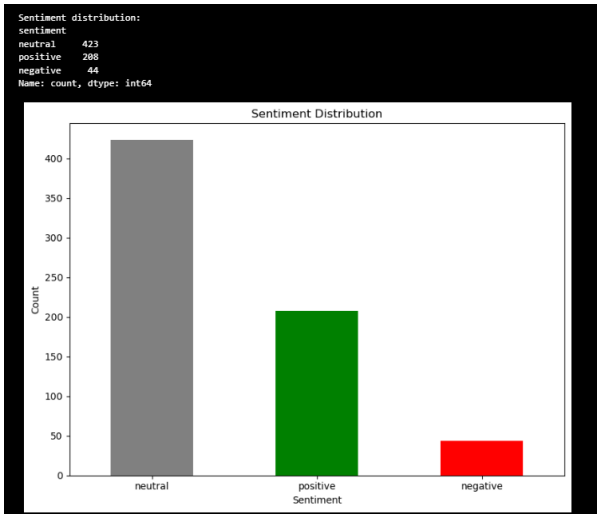


Figure 1: Sentiment Distribution

2. **Word Frequency Analysis**: Identification of the most common terms after stopword removal

3. **Word Cloud Visualization**: Visual representation of term frequency in the corpus

Figure 1:Word Cloud

4. **Statistical Summaries**: Distribution of sentiment categories across the dataset

The visualization components effectively communicate the sentiment distribution of content related to the Netherlands.

Machine Learning Implementation

The project implements a comprehensive model evaluation strategy:

1. **Feature Engineering**: TF-IDF vectorization to convert text to numerical features
2. **Class Imbalance Handling**: SMOTE application to balance the positive and negative sentiment classes
3. **Model Training and Comparison**: Six different classification algorithms evaluated:
    - Logistic Regression - 0.9020
    - Decision Tree – 0.7451
    - Random Forest – 0.9020
    - Gradient Boosting – 0.8854
    - Naive Bayes – 0.7843
    - K-Nearest Neighbors – 0.0980
4. **Model Selection**: Logistic Regression selected as the deployment model based on performance metrics – 0.9020

This thorough comparison demonstrates good machine learning practice by evaluating multiple modeling approaches.

Deployment

The project culminates in a Streamlit web application that:

- Provides a user-friendly interface for sentiment prediction
- Displays prediction confidence scores
- Compares model predictions with TextBlob's default sentiment analysis
- Includes explanatory context about the model's development

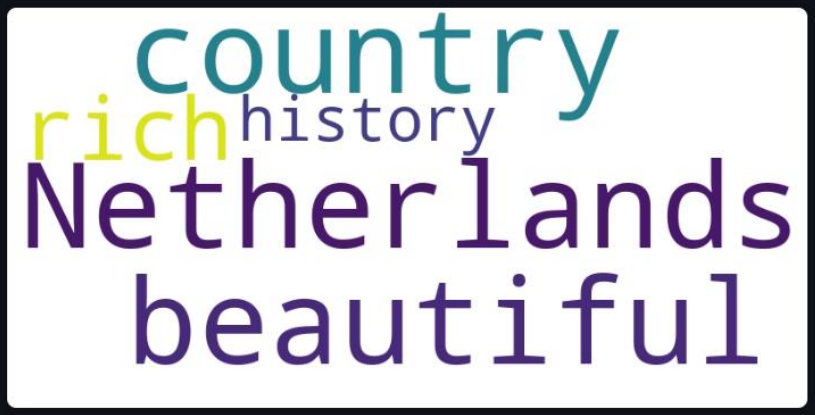
The deployment effectively showcases the project's practical application and makes the sentiment analysis accessible to users.

Netherlands Wikipedia Sentiment Analysis by Malde Saicharan

This app predicts the sentiment of text based on a model trained on Netherlands Wikipedia data. Enter a sentence below to analyze its sentiment.

Enter a sentence:

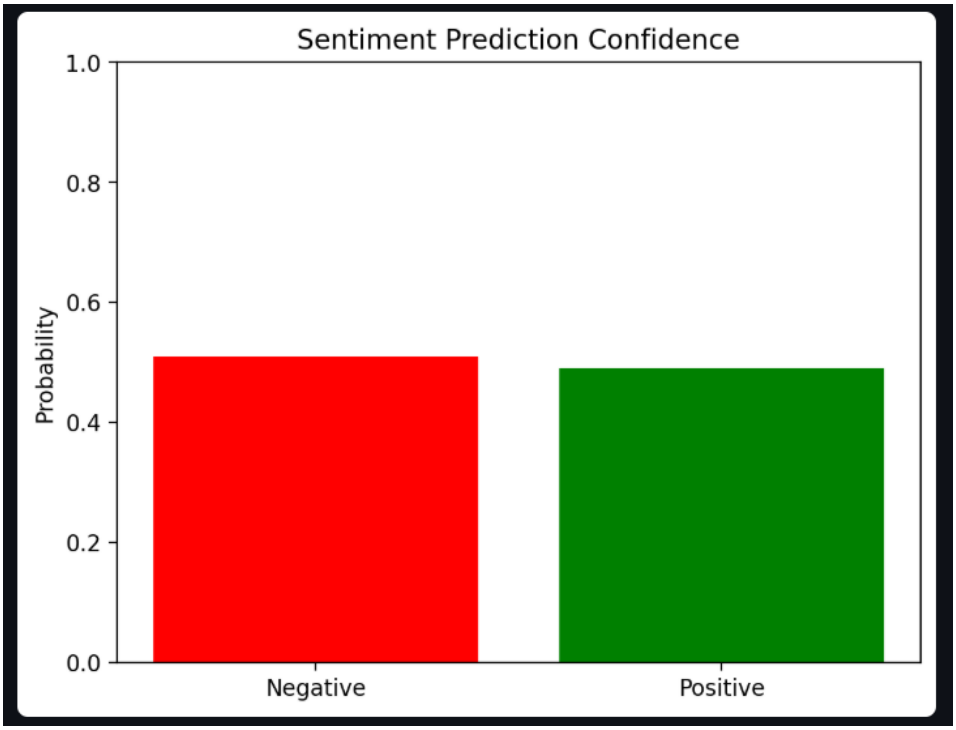
The Netherlands is a beautiful country with rich history.



Word Cloud

Predict Sentiment

Sentiment: Negative



Strengths

1. **Comprehensive Pipeline**: The project demonstrates a complete end-to-end data science workflow.
2. **Methodical Approach**: Systematic implementation of preprocessing, EDA, modeling, and deployment.
3. **Model Comparison**: Thorough evaluation of multiple classification algorithms.
4. **Practical Application**: Interactive web application that makes the model accessible and usable.
5. **Code Organization**: Well-structured code with logical separation of data processing, analysis, and deployment components.

Areas for Enhancement

1. **Sentiment Classification Threshold**: The current sentiment classification uses fixed thresholds (±0.1). Consider exploring optimal threshold selection through ROC curve analysis.
2. **Feature Engineering Expansion**: Additional features like n-grams, POS tags, or entity recognition could potentially improve model performance.
3. **Model Explainability**: Adding SHAP or LIME analysis would provide insight into the model's decision-making process.
4. **Cross-Validation**: Implementing k-fold cross-validation would produce more robust performance metrics.
5. **Error Analysis**: Examining misclassified examples could reveal patterns for model improvement.

Technical Performance

Based on the code shown, the project demonstrates strong technical implementation:

- **Data Collection**: Effective web scraping with status code checking
- **Preprocessing**: Comprehensive text cleaning and normalization
- **Modeling**: Systematic evaluation of multiple algorithms with performance reporting
- **Imbalance Handling**: Proper application of SMOTE to address class imbalance

- **Deployment**: Functional Streamlit application with model serving capability

Conclusion

This Netherlands Wikipedia Sentiment Analysis project demonstrates strong proficiency in NLP, machine learning, and data science workflows. The comprehensive implementation from data collection through deployment shows an excellent understanding of the field's best practices. The Stream lit application effectively showcases the practical application of the sentiment analysis model.

The project succeeds in developing a sentiment analysis system specific to Netherlands-related content with an accessible interface. With the suggested enhancements, particularly around model validation and explainability, this project could be further strengthened for production applications or academic presentation.

Links of Project

saicharan0623/Netherlands-wikipedia-sentimental-analysis

Streamlit