Code Documentation

437 Group:Jean Dong Cho,Samuel Zhang,Ryan Luder,Charles Allbritton

Project Overview

This code constitutes a machine learning project aimed at analyzing political tweets from the Republican and Democratic parties. The primary objectives include exploring language differences between supporters of different political parties, classifying tweets into their respective parties using Natural Language Processing (NLP) techniques, and gaining insights into political discourse on social media.

Architecture

The project is structured as follows:

1. **Data Acquisition and Preparation:**
   - The data is sourced from Kaggle, a reputable machine learning dataset platform, and is stored in a CSV file.
   - Data cleansing is a crucial step, involving the handling of missing values, removal of duplicates, and addressing outliers. Ethical considerations are taken into account during this process.
2. **Natural Language Processing (NLP) Techniques:**
   - The NLTK library in Python is utilized for text preprocessing.
   - Bag-of-Words and Term-Frequency-Inverse-Document-Frequency (TF-IDF) representations are implemented to analyze word frequencies and relationships in the tweets.
   - A custom Neural Network Classifier is developed for party classification, specifying the architecture, activation functions, and regularization techniques.
3. **Data Visualization:**
   - The project includes visualization techniques such as scatter plots to effectively communicate results.
   - Matplotlib and Plotly libraries are used for creating visualizations.
4. **Machine Learning Model:**
   - Scikit-learn's **MultinomialNB** is employed for initial machine learning classification.
   - A custom Multinomial Naive Bayes classifier is implemented.

Design Principles

1. **Modularity:**
   - The code is designed with modularity in mind, allowing for easy integration of additional features or changes in the future.

- Functions are employed to encapsulate specific tasks such as data cleansing, tweet cleaning, and visualization.

2. **Code Reusability:**
   - Commonly used functionalities, such as text cleaning and data visualization, are encapsulated in functions to promote code reusability.
   - The use of functions makes it easier to replicate the code for similar projects or extend functionality.
3. **Ethical Considerations:**
   - Privacy and ethical considerations are addressed during data preprocessing, including anonymization of data and the removal of personally identifiable information.

Implementation Guidelines

1. **Prerequisites:**
   - Ensure that necessary libraries, including Pandas, NumPy, Scikit-learn, NLTK, Matplotlib, and Plotly, are installed in the Python environment.
2. **Data Loading:**
   - Load the dataset into the **/content/drive/MyDrive/** directory or adjust the file path in the code accordingly.
3. **NLTK Downloads:**
   - Download NLTK resources for tokenization and lemmatization using the provided **nlp.download()** commands.
4. **Code Execution:**
   - Execute the code cells sequentially in a Google Colab environment.

Future Directions

1. **Model Fine-Tuning:**
   - Experiment with different hyperparameters for the machine learning model for potential improvements.
   - Consider exploring more advanced neural network architectures for party classification.
2. **Deployment:**
   - Investigate options for deploying the trained model for real-time tweet classification or integration into other applications.
3. **Continuous Monitoring:**
   - Establish a system for continuous monitoring and retraining of the model to adapt to changing language patterns on social media platforms.