Report on Sentiment Analysis for a Twitter Dataset

Objective:

To perform the sentiment analysis of a given dataset containing Twitter sentiments using machine learning and natural language processing.

Motivation:

These days, people express their views about different topics and products on social media. The different social media include Facebook, Twitter, etc. The expressed sentiments can be useful for some researchers and businessmen so that they can get the public views about their articles and products respectively.

Introduction:

Sentiment analysis is a technique used to determine the emotional tone of a piece of text. It is widely used in various fields such as marketing, customer service, and politics. Data preprocessing is an important step in sentiment analysis for converting the text into the form of words with just spaces between them. Also, the stemming process is required for converting the words into their root forms. Moreover, the sentiments are classified based on the polarity and labels.

Overview of the Project:

This program first performs the analysis of data and drops the unnecessary columns as only ‘text’ is to be dealt with. Also, the different functions like ‘data_preprocessing’, ‘sentiment’, ‘polarity’, and stemming are defined for cleaning, classification of sentiments, polarity determination, and stemming the words respectively.

Methodology

a. **Importing Libraries and Modules:**

The first step in data preprocessing is to import the necessary libraries and modules. In this guide, we use the following libraries: nltk, textblob, pandas, numpy, re, seaborn, matplotlib, and wordcloud. These libraries provide various functions and tools for data preprocessing, visualization, and analysis.

b. **Dropping Unnecessary Columns:**

The next step is to import the dataset and drop any unnecessary columns. In this guide, we use a dataset of tweets and drop the 'selected_text' and 'sentiment' columns. This is because we only need the 'text' column for sentiment analysis.

c. Data Preprocessing:

After dropping unnecessary columns, data preprocessing is conducted using NLTK and TextBlob. The data preprocessing techniques include

- Converting text to lowercase
- Removing URLs and mentions
- Removing special characters and punctuation
- Tokenizing text into words –
- Removing stop words
- Removing the numeric values

d. Stemming and Polarity determination:

Once the data is preprocessed, stemming is performed using the PorterStemmer function. This reduces words to their root form, which helps in reducing the number of unique words in the dataset. The polarity of each tweet is calculated using the TextBlob function. Polarity is a measure of the sentiment of a piece of text, ranging from -1 (negative) to 1 (positive). Sentiment Analysis: After calculating the polarity, a function is defined to determine the sentiment of each tweet based on its polarity. Then, the distribution of sentiments is plotted using a count plot and a pie chart. Also, the tweets are sorted in descending order for all cases (negative, positive, and neutral) based on obtained polarity and the word clouds are created for the most positive and negative tweets.

Discussion

This program successfully classifies the sentiment into positive, negative, and neutral based on the polarity and then presents them in the form of a bar graph and pie chart. The accuracy is above 90% which indicates that our model (Logistic Regression) is appropriate for this classification problem. Also, hyperparameter tuning is done using ‘GridSearchCV’ by providing a list of different parameters and getting the best one.

Conclusion:

This program can be concluded as a successful demonstration of sentiment analysis using the concepts of Machine learning, Data Analysis, and Natural Language processing. The hyperparameter tuning helps in the improvement of accuracy. Further, the confusion matrix and classification reports indicate the better performance of the model. In the future, more hyperparameters can be used for the better performance of the model.

References:

-https://jovian.com/learn/data-analysis-with-python-zero-to-pandas

-https://jovian.com/learn/machine-learning-with-python-zero-to-gbms

-https://www.youtube.com/watch?v=RLfUyn3HoaE&t=1889s

-https://www.youtube.com/watch?v=ng6L_wvREB4&t=1s

-https://chat.openai.com

Project Links:

https://github.com/AsmitaJha/Sentiment-Analysis-on-Dataset/blob/main/README.md

https://jovian.com/asmitajha174/nlp1?username=asmitajha174&params=nlp1&jobot=open