Summary

Title

Sentiment Analysis Using MNB, MLP, and SVM

Date

October 2023

Project Objective

- To compare the performance of three classification algorithms Multinomial Naive Bayes (MNB), Multilayer Perceptron (MLP), and Support Vector Machine (SVM)
- Compare and contrast the performance of these models using both preprocessed and raw data, and then explain findings.

About Dataset

Dataset: tweet_emotions.csv
Collected from data.world platform, shared by @crowdflower, a data enrichment, mining and crowdsourcing company based in the US.
https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text

Data Features

- Dataset consists of 40000 records of tweets labelled with 13 different sentiments, followed by Tweet ID number.
- All data types are strings, except for the 'tweet ID' column, which is an integer.
- There is a class imbalance of <21.32%. Tweets primarily convey neutral and negative sentiments.
- Contains 172 duplicate rows (tweets) but categorised with different sentiment labels.
- Word lengths vary, ranging from a minimum of 1 word to a maximum of 16 words.
- The data exhibits some degree of disorder and lack of cohesion. Various linguistic patterns are used to express sentiment.
- Some information contains only special characters or hyperlinks.
- Contains informal and colloquial terms. For instance, “peeps” is a friendly term for “People”.
- Contains self-made terms, slangs or misspellings such as “Humpalow”.

What I Did to Mitigate Data Imbalance

Combined the 13 labels into three primary emotions: Positive, Negative, and Neutral.

4 Preprocessing Steps I Made to Maximise Accuracy

1. Cleaned special characters
2. Lemmatisation
3. Retained stopword
4. TF-IDF as Feature Engineering method

Model 1: MNB:

Accuracy: 53.3% Parameters: default

Model 2: MLP

Accuracy: 53%, Parameters: 100,100, a=0.01

Model 3: SVM

Accuracy: 59%, Parameters: default

Best Model and reason

Decision Tree Classifier (DTC), because:

- SVM demonstrated the highest Accuracy, Precision and Recall rates.
- It exhibits robustness to noise and changes in input data
- Faster to compute

Tools used

Python (Natural language processing (NLP) libraries like NLTK, and Machine Learning libraries like scikit-learn).