News Sentiment Analysis System

Project Description

News sentiment analysis is a system that automatically identifies the sentiment (positive, neutral, or negative) of news articles or headlines. The purpose of this project is to develop a machine learning algorithm for news sentiment analysis using a given dataset. The dataset consists of news articles and their associated sentiment labels, and the goal is to train a classifier that can accurately predict the sentiment of new articles.

- The number of classes in the dataset is three (positive, neutral, and negative).
- The dataset consists of 4846 news articles
- 60% of the data for training and 40% for testing

Methodology

- Data-preprocessing
  - Giving dataset column heads -> sentiment, text
  - removing URLs, special characters, and stop words.
  - stemming the words using the PorterStemmer algorithm
- split the preprocessed data into training and testing sets using a 60/40 split.
- used the TF-IDF vectorizer to convert the preprocessed text data into numerical feature vectors
- Classifiers
  - Multinomial Naive Bayes
  - Gaussian Naive Bayes

Results

The classification reports for the Multinomial Naive Bayes and Gaussian Naive Bayes classifiers are as follows:

Multinomial Naive Bayes classification report:

Class	Precision	Recall	F1-score	Support
negative	0.82	0.06	0.11	231
neutral	0.68	0.97	0.80	1149
positive	0.68	0.34	0.45	559
Accuracy			0.68	1939
Macro avg	0.73	0.46	0.45	1939
Weighted avg	0.70	0.68	0.62	1939

Gaussian Naive Bayes classification report:

Class	Precision	Recall	F1-score	Support
negative	0.31	0.50	0.39	231
neutral	0.70	0.56	0.62	1149
positive	0.36	0.42	0.39	559
Accuracy			0.51	1939
Macro avg	0.46	0.50	0.47	1939
Weighted avg	0.56	0.51	0.53	1939

Based on the classification reports, we can see that the Multinomial Naive Bayes classifier outperformed the Gaussian Naive Bayes classifier on all metrics. The Multinomial Naive Bayes classifier achieved an overall accuracy of 0.68, with the highest precision and F1-score for the 'neutral' sentiment class (0.68 and 0.80, respectively). However, the classifier performed poorly for the 'negative' and 'positive' classes, with precision and recall scores below 0.70.

One potential problem with making false classifications is that it can lead to misleading information. For example, if a positive news article is misclassified as negative, it could cause unnecessary concern among readers. Similarly, if a negative news article is misclassified as positive, it could lead to lack of caution among readers.

Another potential problem is that misclassifications can impact the performance of downstream applications that rely on the sentiment analysis results. For example, if a stock trading algorithm relies on sentiment analysis to make buy or sell decisions, a misclassification could lead to poor investment decisions and financial losses.