IST 664: Natural Language Processing

Name: Jovita Andrews

SUID: 522110149

**Problem Definition**

The goal of this project is to develop a model capable of categorizing the sentiment expressed in movie reviews into five categories: very negative, negative, neutral, positive, and very positive. This task is significant as it helps in understanding customer perceptions and can be utilized for services such as recommendation systems or market analysis.

**Data Preparation**

The dataset was obtained from Kaggle and consists of movie reviews split into two files: training and test datasets. The training set includes 156,060 entries with features such as PhraseId, SentenceId, Phrase, and Sentiment, while the test set contains 66,292 entries missing the Sentiment labels, which are to be predicted.

**Data Loading:** Data from a zipped archive was extracted and loaded using Pandas. The `train.tsv` and `test.tsv` were loaded with fields separated by tabs.

**Data Inspection**: The training data contains no missing values, and sentiment distribution shows that most phrases are neutral, followed by somewhat positive and somewhat negative sentiments. A single null value in the test set needs to be handled before modeling.

**Feature Engineering:** The maximum length of phrases is determined (52 words), and new features such as sentiment phrase labels and word counts are added for better analysis.

**Exploratory Data Analysis:**

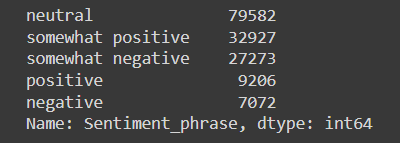
1. **Data Structure:**

A screen shot of a computer

Description automatically generated

* *PhraseId:* Unique identifier for each phrase.
* *SentenceId:* Identifier linking phrases from the same sentence.
* *Phrase:* Text content of the phrase.
* *Sentiment:* Numerical label indicating the sentiment of the phrase, ranging from 0 (very negative) to 4 (very positive).

1. **Sentiment Distribution:**



The distribution of sentiment labels is uneven, with middle values (presumably neutral sentiments) being more frequent.

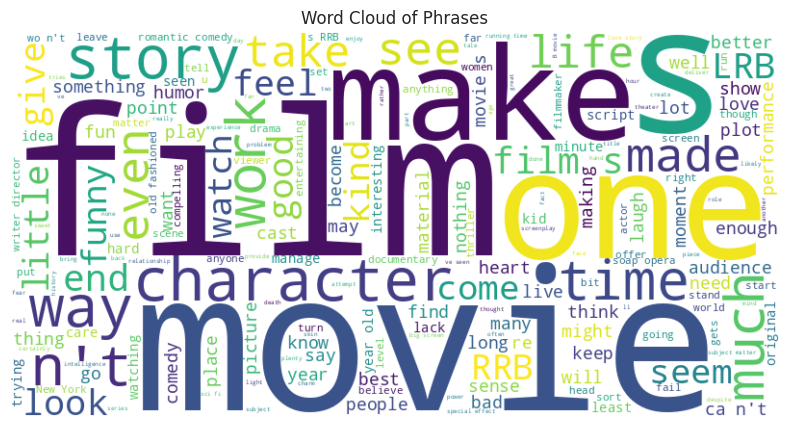
1. **Phrases per Sentence:**

**A graph of a number of phrases

Description automatically generated**

Most sentences contain only a few phrases, suggesting concise breakdowns within sentences. Maximum number of words found in a particular sentence is ‘52’

1. **Word Cloud Visualization:**

****

Frequent words in the dataset include terms like 'film', 'movie', 'story', and 'like', which are common in movie reviews.

1. **N-Gram Analysis:**

**A screenshot of a computer screen

Description automatically generated**

* Bi-grams: Common bi-grams include 'romantic comedy', 'year old', and 'new york'.
* Tri-grams: Notable tri-grams reflect specific themes or phrases like 'world war ii' and 'big fat greek'.

1. **Example Sentences for Sentiment Categories:** A screenshot of a computer

   Description automatically generated

* Negative: Examples indicate criticisms or negative aspects noted by reviewers.
* Neutral: Typically, factual or neutral descriptions.
* Positive: Praise or positive comments about movies, indicating enjoyment or admiration.

**Modeling**

For the sentiment analysis model, both traditional machine learning and deep learning approaches were considered:

**Naive Bayes Classifier:**

**A number on a black background

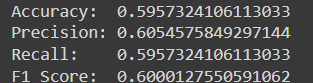
Description automatically generated**

*Setup and Training:* A Multinomial Naive Bayes classifier was used due to its suitability for handling frequency-based vector representations of text data. The model was trained using the TF-IDF vectorized training data.

*Performance on Training Data:* The model achieved an accuracy of approximately 60.8% on the training set, indicating a reasonable fit to the dataset.

*Visualization:* A heatmap of the confusion matrix was created to visually assess the model's performance in correctly predicting each sentiment class.

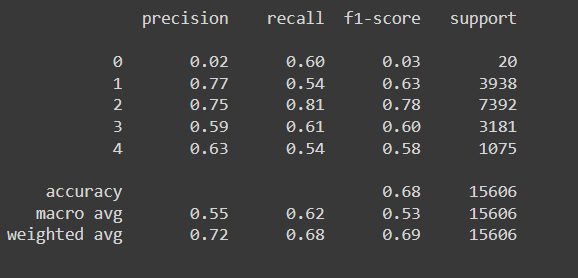
**Decision Tree Classifier:**

****

*Setup and Training:* A Decision Tree Classifier was also implemented as a comparative model to understand the potential of simple decision-based algorithms in handling textual data.

*Performance Evaluation:* The Decision Tree model slightly underperformed compared to the Naive Bayes, with an accuracy of 59.5%, precision of 60.5%, recall of 59.5%, and F1 score of 60% on the validation set.

**Bidirectional Encoder Representations from Transformers (BERT):**



The sentiment analysis was further enhanced by leveraging the BERT model, known for its state-of-the-art performance in various NLP tasks. Here's a detailed look at the setup and observations from using BERT.

**Model Setup:**

* *Model Configuration:* We employed bert-base-uncased which has 109 million parameters. BERT's architecture uses transformer layers, which enable it to capture the context from both left and right sides of a token within a sentence.
* *Input Layer Configuration:* The input sequences were tokenized and prepared with a maximum length of 45 tokens to ensure consistency across all data points and to optimize computational efficiency.

**Model Training:**

* *Compilation:* The model was compiled with an Adam optimizer, categorical cross-entropy loss, and tracked accuracy as a metric. This setup is suitable for multi-class classification tasks like sentiment analysis.
* *Training:* The model was trained for a relatively short period (2 epochs) to prevent overfitting, given BERT's capability to quickly learn from vast amounts of text data.

**Model Performance and Evaluation:**

* *Accuracy and Loss:* BERT achieved an accuracy of approximately 68% on the validation set. While not groundbreaking, this accuracy is substantial given the complexity of sentiment classification in a multi-class setup.
* *Precision, Recall, and F1-Score:* The model demonstrated a weighted average precision of 72%, recall of 68%, and an F1-score of 69%. The highest recall and precision were observed for the 'neutral' class, which is often the most prevalent in sentiment datasets.
* *Confusion Matrix Analysis:* The displayed confusion matrix showed that while the model performs well in recognizing neutral sentiments, it struggles with extreme sentiments (very negative and very positive), which are less represented in the data.

**Predictions on Test Data:**

* *Test Data Predictions:* The model predictions on the test data suggested that BERT effectively generalizes from the training data to unseen data. The softmax outputs indicate the model's confidence in each class, showing robustness in its predictions.
* *Further Metrics Computation:* Advanced metrics such as average precision score, weighted F1 score, and recall were computed, demonstrating BERT's nuanced understanding of class analysis, making it a preferred choice for high-stakes applications where accuracy is critical.

**Conclusion:**

The project aimed at developing a model to classify movie review sentiments into five distinct categories achieved considerable success by implementing and comparing multiple approaches. The Multinomial Naive Bayes model, leveraging TF-IDF vectorization, provided a decent baseline accuracy of around 60.8%, while the Decision Tree Classifier slightly underperformed but still offered valuable insights with a 59.5% accuracy. The use of the BERT model marked a significant improvement, achieving around 68% accuracy on the validation set, demonstrating its superior ability to capture complex linguistic nuances across a variety of sentiments, especially in neutral contexts. Despite challenges in accurately predicting extreme sentiments, BERT's advanced capabilities in handling contextual information and its robust performance on unseen test data highlight its potential for practical applications in sentiment analysis, making it a promising tool for enhancing recommendation systems or market analysis through nuanced understanding of customer feedback.

**Insights and Future Directions:**

The Naive Bayes and Decision Tree classifiers provided a foundational understanding of the dataset's complexity and the challenges in sentiment classification. The LSTM model's potential was acknowledged but requires further implementation and tuning. Going forward, enhancing the LSTM model, exploring more sophisticated architectures like GPT-4 for contextual embeddings, and extensive hyperparameter tuning are critical steps. Additionally, expanding the feature set and employing techniques like cross-validation could further improve the robustness and accuracy of the predictive models.