Topic Name: Sentiment Analysis for Financial News

Team Member Details:

Sneh Patel – C0869367
Malav Shah- C0870066
Dhruvraj Chavda – C0867457
Nandiniben Patel – C0869773

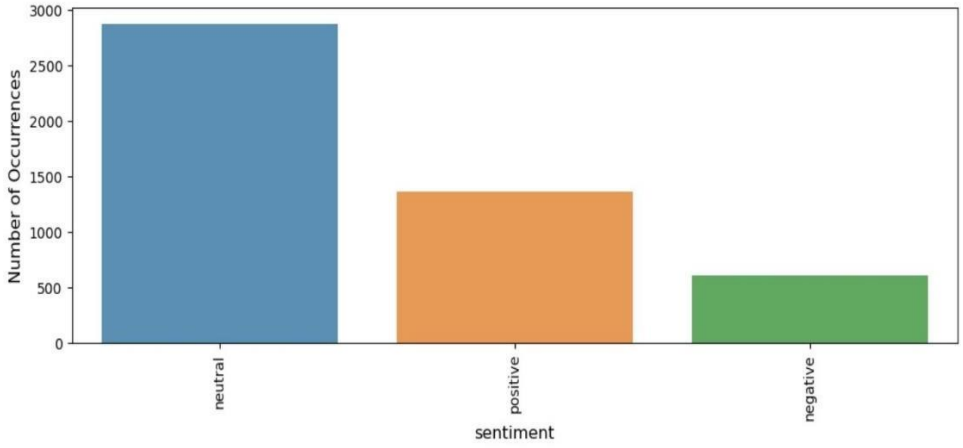
Introduction:

- Financial news items and the related attitudes from the viewpoint of a retail investor are included in the dataset FinancialPhraseBank. The dataset has two columns: one for the news headline and one for the sentiment, which might be favorable, neutral, or unfavorable.

- The Financial Phrase Bank dataset may be used for a variety of natural language processing applications, including sentiment analysis, text categorization, and topic modeling.

Dataset Description:

- The dataset description on Kaggle contains over 4,845 financial news headlines labeled with 1,363 positive, 2,878 neutral and 604 negative sentiments categories.



- The data stored in CSV file format has columns including "Sentiment" (the sentiment label) and "Message" (the financial news headline).

- The emotion label, which can be neutral, positive, or negative, expresses how a retail investor feels about the news headline. The news headlines, compiled from various sources, cover a wide range of financial issues, including stock prices, economic indicators, mergers and acquisitions, and corporate results.

- The dataset is relatively balanced, with each sentiment class making up around one-third of the data. The news headlines range from 2 to 22 words, with an average size of about ten words.

Data Cleaning and Text Preprocessing Explanation:

- Data cleaning and text preparation must be done prior to training machine learning models on the Financial Phrase Bank dataset. Any useless or contradictory information, such as duplicate entries or missing values, may be eliminated at this stage.

- Moreover, any unique characters or symbols that can obstruct the research should be eliminated.

- To prepare the data for analysis, several operations like tokenization, stop word removal, stemming or lemmatization, and feature extraction might be carried out. The text preparation methods selected rely on the study's unique requirements.

- Term TF-IDF is a quantitative measure of a term's significance to a text in a corpus. It is computed by dividing the inverse document frequency by the term frequency or how many times a term is used in a document. The resultant value gives words specific to a given text more weight and terms standard throughout the corpus less weight.

- The Bag of Words approach is used to represent text data as a group of words that aren't necessarily in any particular order or sequence. It entails developing a corpus-wide vocabulary of distinct terms and then encoding each text as a frequency vector of the language's words. The sequence and context of the words are lost, but it still enables us to assess the usage and frequency of each term in a document.

- A bigram is a grouping of two similar words that appear in a text. Some of the contextual information that is lost in the Bag of Words model can be captured by bigrams. We may establish a vocabulary of all potential bigrams in the corpus and then express each document as a frequency vector of the bigrams to generate a bigram representation of the text data.

- Similarly, a trigram is a grouping of three words that appear in a row together in writing. Trigrams have the potential to capture even more information than bigrams, but they may also produce more features and take longer to analyze. We may build a vocabulary of every conceivable trigram in the corpus and then describe each document as a frequency vector of the trigrams to generate a trigram representation of the text data.

Model Building Process and Results Interpretation and Comparison:

- For the Financial Phrase Bank dataset, various machine learning methods may be utilized to develop a sentiment analysis model, including logistic regression, support vector machines, decision trees, and neural networks.

- When the models have been trained, the outcomes may be assessed using metrics like accuracy, precision, recall, and F1 score.

- The exact objectives of the study will determine the assessment measures to be used. The models can also be contrasted using cross-validation methods or based on their performance on a validation dataset.

- Afterward, the model deployment with the best performance may be chosen.

- For the purpose of creating automated sentiment analysis machine learning models, the Financial Phrase Bank dataset might be helpful.

- The dataset can yield valuable insights that can be utilised to guide decision-making in a variety of financial scenarios by undertaking data cleaning and text preparation, as well as creating and assessing machine learning models.

Training & Testing:

- In machine learning, training and testing datasets are frequently used to assess a model's performance. The model is trained using the training dataset, and its effectiveness is assessed using the testing dataset.

- A portion of the data used to train the model is called the training dataset. It frequently has a lot of instances, each having a well-known input and outcome. On the basis of this data, the model is trained by changing its parameters to reduce the discrepancy between the output it predicts and the actual output.

- The performance of the trained model is assessed using the testing dataset, a different subset of the data. It's crucial to avoid using the testing data during training since doing so might result in overfitting, which causes the model to perform well on

training data but badly on brand-new, untested data. The testing data should be an accurate representation of the actual data that will be used to test the model.

- We perform Training and Testing on the model, containing 3876 training datasets and 969 testing datasets.

Conclusion:

- These accuracy scores indicate the percentage of correctly classified instances by the three models - Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine.

- The Multinomial Naive Bayes model achieved an accuracy score of 71.10%, meaning that 71.10% of the instances were correctly classified.

- The Logistic Regression model achieved an accuracy score of 75.54%,,slightly better than the Naive Bayes model.

- The Support Vector Machine model achieved an accuracy score of 69.56%, slightly worse than the Naive Bayes model.

- The higher accuracy score for Logistic Regression indicates that it may be the best model for this dataset.

Based on the given accuracy scores, we can conclude that the logistic regression model performs the best among the three models tested.

References

Sentiment Analysis for Financial News. (2020, May 27). Kaggle.
https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news