Project Report: Sentiment Analysis on Financial News

Introduction:

Sentiment analysis is a natural language processing (NLP) technique that involves determining the sentiment expressed in a piece of text. In the financial domain, sentiment analysis on news articles can be crucial for understanding market trends, investor sentiment, and potential impacts on stock prices. This project focuses on building and evaluating a sentiment analysis model for financial news using both a custom-built LSTM model and a pre-trained language model. This project was a part of the CFA Level 2 Machine learning and Artificial Intelligence module.

Dataset:

The dataset includes text data and corresponding labels indicating the sentiment of each news article (positive, negative, or neutral).

Methodology:

- **Data Exploration**
  - The initial exploration of the dataset involves examining its structure, displaying the first few rows, and checking the distribution of sentiment labels. The dataset contains a mixture of positive, negative, and neutral sentiment samples.

- **Data Cleaning**
  - Before feeding the text data into the models, a data cleaning process is implemented. This involves removing non-alphanumeric characters, converting text to lowercase, and removing stop words. The cleaned text is then added as a new column in the dataset.

- **Word Cloud Visualization**
  - Word cloud visualizations are generated for each sentiment category (positive, negative, and neutral) to gain insights into the most frequent words used in each sentiment class. Custom stop words related to politics and stocks are excluded from the word clouds.

- **Model Building - LSTM**
  - ***Tokenization and Padding***
    - The text data is tokenized using the FinBERT tokenizer, and padding is applied to ensure consistent input dimensions for the LSTM model.

    - ***Model Architecture***
        - An LSTM-based neural network is built using the Keras library with TensorFlow backend. The model includes an embedding layer, an LSTM layer with 64 units, and a dense layer with softmax activation for multi-class classification. The model is compiled using the Adam optimizer and sparse categorical cross entropy loss.

    - ***Model Training***
        - The LSTM model is trained on the training dataset, and its performance is evaluated on the validation set. The training history is recorded for later analysis.

- **Pre-trained Language Model - FinBERT**
    - A pre-trained language model (FinBERT) is utilized for sentiment analysis without fine-tuning. The model is applied to predict sentiment labels for the testing dataset, and the accuracy of the pre-trained model is compared with the custom-built LSTM model.

Results and Discussion:

- **LSTM Model Results:**

  The LSTM model achieved an overall accuracy of approximately 75% on the test dataset. The classification report indicates variations in precision, recall, and F1-score across the three sentiment classes. The model performed better in predicting neutral sentiments compared to positive and negative sentiments.

- **Pre-trained Model Results:**

  The pre-trained FinBERT model, without fine-tuning, achieved an accuracy of approximately 71.72% on the testing dataset. This performance is competitive with the custom-built LSTM model, showcasing the effectiveness of pre-trained language models for sentiment analysis tasks.

Future Work:

- Fine-tune the pre-trained models on the specific financial news dataset to potentially improve performance.
- Explore other transformer-based architectures for sentiment analysis, such as *BERT* and *GPT*-based models.
- Investigate the impact of hyperparameter tuning on the performance of the custom-built LSTM model.
- Enhance the interpretability of the model's predictions by analyzing misclassifications and incorporating attention mechanisms.