Proposal for “Exploring FinBERT for Stock News Sentiment Analysis: A Comparative Project with LSTM”

Yoona Lee
Master’s in Data Science
University of Michigan
Ann Arbor, Michigan, USA
yoonalee@umich.edu

I. OVERVIEW

In the financial domain, timely and accurate sentiment analysis of stock-related news is crucial for informed decision-making. By analyzing sentiment, investors can anticipate positive or negative trends, providing a strategic edge in portfolio management. A positive sentiment often signals potential upward market trends, while a negative sentiment can indicate impending declines. These sentiment shifts offer valuable insights into how market participants perceive the potential impact of news on a company's stock price, enabling predictions about market behavior with greater precision.

While reading news articles, I noticed comments about economic trends, such as those affecting the stock market, were often removed. This led me to wonder if such comments might negatively impact certain companies. Manually reviewing articles and identifying harmful comments would be labor-intensive and costly. This inspired the idea of using web crawling to scrape news articles and develop a system to filter potentially harmful comments efficiently. After learning about deep learning and classification techniques in class, I decided to explore this topic further through analysis and modeling.

This project focuses on FinBERT, a pre-trained language model specifically optimized for financial sentiment analysis, as the primary tool for analyzing stock-related news sentiment. To evaluate its effectiveness, we compare FinBERT's performance with Long Short-Term Memory (LSTM) networks, a widely used RNN architecture for sequential data processing. Additionally, we explore the potential of integrating FinBERT-generated sentiment embeddings as input features for the LSTM model to assess whether this hybrid approach can yield improved results.

The proposed methodology comprises three key stages: (1) applying the pre-trained FinBERT model to classify the sentiment of fiance(stock-related) news, (2) independently training an LSTM network to capture temporal patterns and contextual dependencies within the same dataset, and (3) integrating FinBERT-generated sentiment embeddings as additional input features for the LSTM model. This framework highlights FinBERT’s role as the foundation of the analysis, while also examining how it complements traditional RNN-based approaches. By combining FinBERT's robust language understanding capabilities with LSTM's ability to model sequential relationships, we aim to enhance sentiment analysis performance beyond what either model can achieve individually.

The efficacy of this approach is evaluated using benchmark datasets, with performance metrics such as accuracy, precision, recall, and F1-score. Comparative experiments demonstrate the strengths and limitations of FinBERT and LSTM models individually, as well as the potential performance gains from their integration. This investigation not only reinforces FinBERT's suitability for financial sentiment analysis but also explores the practicality and advantages of integrating transformer-based and RNN-based models for enhanced performance in this domain.

Through this work, I aim to advance the state-of-the-art in financial text analysis by emphasizing FinBERT's capabilities and exploring its integration with complementary techniques.

II. PRIOR WORK

One significant work reviewed was *"FinBERT: Financial Sentiment Analysis with Pre-trained Language Models"*[1], which provided valuable insights into the application of domain-specific pre-trained language models for sentiment classification in financial texts. This paper highlighted FinBERT's ability to achieve state-of-the-art performance on datasets such as the Financial PhraseBank, emphasizing its effectiveness in addressing the unique linguistic challenges of financial text.

Predicting stock market movements has been a challenge in the financial domain. Over time, researchers have increasingly relied on sophisticated machine learning and deep learning methodologies to improve predictive accuracy. Traditional methods that utilized technical indicators derived from stock price data have evolved into more comprehensive approaches that incorporate textual data, such as financial news and social media (Twitter, Meta) sentiment. This integration aims to capture market sentiment, which plays a crucial role in stock price fluctuations.

Early approaches in this field combined machine learning with lexicon-based sentiment analysis but struggled to capture the complexity of financial text. The introduction of deep learning models, especially LSTMs, revolutionized the field by effectively processing sequential data, making them ideal for tasks like stock market prediction. However, RNNs, including LSTMs, suffer from limitations such as limited contextual understanding within the text.

Another approach, ELMo (Embeddings from Language Models), is a deep learning technique that overcomes the limitations of traditional word-embedding methods by learning contextualized word embeddings through a bi-directional language model. Unlike traditional approaches, ELMo captures the context-dependent meaning of words, enabling it to represent word relationships and enhance performance in NLP tasks effectively.

Extending this idea, recent stock market forecasting models have incorporated attention mechanisms, such as self-attention and transformers. Transformers, like BERT, overcome the limitations of RNNs and ELMo by providing fully bidirectional contextual representations across all layers. FinBERT, a BERT model fine-tuned for financial text, has excelled in sentiment analysis, achieving advanced results in tasks like sentiment classification. This captures the unique vocabulary and semantic structures of financial text, enabling more accurate sentiment classification.

By combining FinBERT's contextual understanding of financial text with LSTM's sequential data processing capabilities, this project aims to enhance the precision of sentiment analysis by fine-tuning and comparative analysis.

III. PRELIMINARY RESULTS

The dataset used for this project consists of 142,000 rows and 9 columns, containing stock news data. A new DataFrame was created by selecting only the variables relevant for modeling. Specifically, a refined dataset was constructed with the following features: title, text, sentiment (classified as negative, neutral, or positive), and sentimentScore. Except for the sentimentScore feature, there were no missing values in the dataset.

Regarding the target variable, sentiment class distribution, the distribution is illustrated in the figure below.

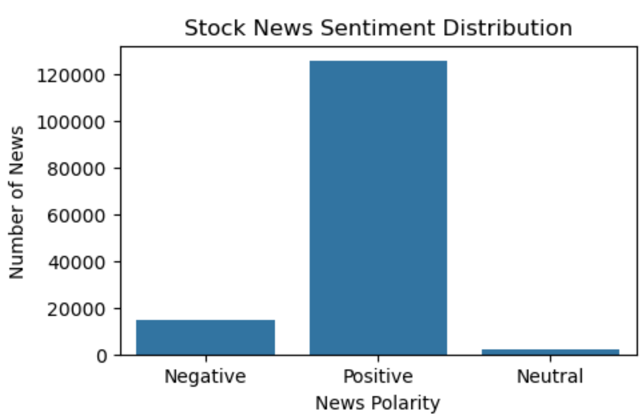

Fig 1. Sentiment Class Distribution

The positive class significantly outweighs the neutral and negative classes, raising concerns about potential overfitting. To address this imbalance, strategies such as oversampling for minority classes, undersampling for the majority positive class, or implementing a specialized metric to mitigate overfitting may be necessary.

To begin, I selected FinBERT as the base model. FinBERT requires a large amount of text data and substantial computing resources to leverage the pre-trained model in NLP tasks, such as sentiment classification, for a much smaller labeled data set, a process known as fine-tuning. Fine-tuning with slanted triangular learning rates and gradual unfreezing can extend the training duration but is essential to prevent catastrophic forgetting. Regarding performance bottlenecks, fine-tuning a pre-trained language model like BERT can cause the model to lose important general-purpose language understanding (catastrophic forgetting) as it adapts to the target task. So we need to apply different learning rates to layers to retain foundational information in lower layers.

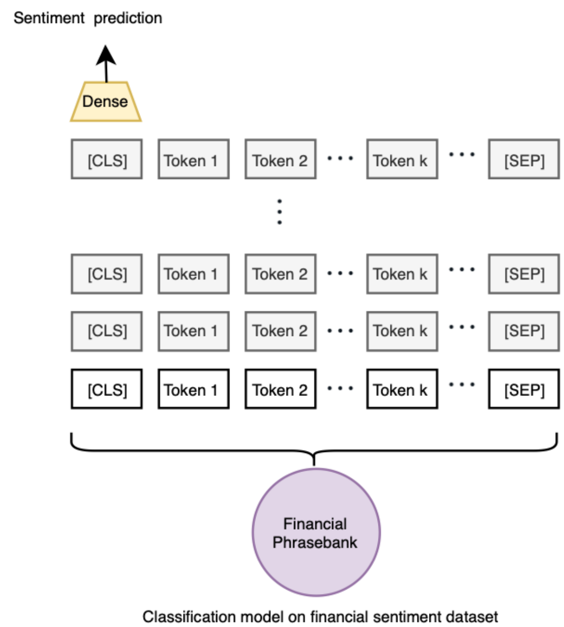

Fig 2. Overview of classification fine-tuning

Another approach, LSTM models are computationally less demanding than transformer-based models like FinBERT. This models train faster than transformers due to fewer parameters and simpler architecture. As previously mentioned in the prior work section, LSTM models have limitations such as limited contextual understanding. Also, LSTM models are heavily dependent on pre-trained static embeddings, which are less adaptable to domain-specific nuances. Therefore, this aspect should be carefully considered during the modeling process.

Techniques such as tokenization, text preprocessing, and classification metrics were applied. To compare performance, FinBERT's pre-trained embeddings and LSTM architecture were also implemented. Further pre-training of FinBERT on financial-specific corpora and parameter tuning for both models will be explored. I will primarily use various tools learned in class, such as Pandas, PyTorch, neural network training, and visualization. Tools like TensorBoard will be used to visualize training progress and identify bottlenecks.

IV. PROJECT DELIVERABLES

This project aims to leverage FinBERT, a pre-trained language model optimized for financial text, to enhance sentiment analysis of stock-related news. By fine-tuning FinBERT , comparing its performance with LSTM models, and integrating its embeddings with LSTM models, the sub-goals in this project are to improve prediction accuracy and address contextual understanding limitations.

In addition, the following deliverables will be produced as part of this project:

1) Datasets: A well-preprocessed and organized finance dataset will be provided for training and testing the model. The dataset will be structured to prioritize usability and thoroughness, supporting future research and development efforts.

2) Explanation about pre-trained models: an overview of the baseline methods used for comparison, emphasizing why FinBERT was chosen as the primary model for this project.

3) Implementation Details: This provides technical details about the models implemented in the project, including the architecture of FinBERT, LSTM, and the hybrid model. It will also include any modifications made to pre-trained models.

4) Performance Evaluation: This section will outline the metrics used to evaluate model performance, such as accuracy, precision, recall, F1-score, and confusion matrices. It will also discuss how they help quantify the effectiveness of FinBERT and LSTM models.

5) Fine-tuning Process: This will include the steps for training hyperparameters, and techniques used to prevent overfitting, and the performance result.

6) Conclusion and Future Work: Summarize the findings of the project, highlighting the key results and contributions. It will also propose directions for future work, such as exploring additional datasets, integrating more advanced architectures, or applying the methodology to other domains.

All of these are included in the source code and the final report, which are the ultimate deliverables of the project.

V. TIMELINE

- Week1.

1) Clearly define the project objectives and expected outcomes

2) Gather the dataset and clean the data

3) Begin the groundwork for the proposal

- Week2.

1) Load the pre-trained model from Hugging Face

2) Fine-tune FinBERT for sentiment classification

3) Assess the performance of the fine-tuned FinBERT model

- Week3.

1) Fix errors encountered during FinBERT fine-tuning

2) Design and build the LSTM model

3) Compare LSTM's performance with FinBERT

4) Finalize and review the proposal

- Week4.

1) Integrate FinBERT with LSTM by extracting sentiment embeddings from the fine-tuned FinBERT model and use this as additional input features for the LSTM Model

2) Train and evaluate the hybrid model

- Week5.

1) Summarize performance comparisons among FinBERT, LSTM, and the hybrid model

2) Complete the final report and develop the source code

REFERENCES

[1] Dogu Tan Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models ”, August 2019.