A Comparison of Sentiment Analysis on Financial Headlines using DistilBERT and LSTM Models

1st James Jung
Department of Statistics
University of Michigan
Ann Arbor, MI, USA
jameshj@umich.edu

***Abstract*—This study investigates the performance of sentiment analysis models on financial headlines, focusing on DistilBERT, a lightweight transformer-based model, and two LSTM models. The project leverages datasets of varying sizes to analyze how model performance changes with the inclusion of additional data. Dataset 1, containing approximately 12,000 rows, serves as the baseline, while Dataset 3, augmented with 9,000 additional rows, offers insights into the impact of dataset expansion. Metrics such as accuracy, F1 score, precision, recall, and loss are used to evaluate model performance.**

Results indicate that DistilBERT consistently outperforms both LSTM models, achieving the highest accuracy and balanced performance across metrics. The Keras-based LSTM demonstrates moderate performance, while the second LSTM lags significantly. Expanding the dataset improves DistilBERT's metrics, except for a slight increase in loss, suggesting its ability to capture more nuanced patterns in larger datasets. The Keras LSTM maintains stable performance, while the second LSTM benefits most from the increased dataset size.

This work highlights the efficacy of pre-trained transformer models for financial sentiment analysis and underscores the importance of dataset diversity and size. Future work could explore advanced models like FinBERT, temporal data integration, and hyperparameter optimization to further enhance sentiment analysis in financial domains.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Sentiment Analysis has been a great boon for Data Scientists as data expanded in scale during the digital age. News headlines and updates have been posted on Twitter to help facilitate the dissemination of information. Finance has been a field of general interest in the application of data science to real-world industries and a foyer into this field is desired.

A significant trend in financial sentiment analysis has been the use of pre-trained transformer models, especially models like BERT and its financial variant, FinBERT, which is pre-trained on financial datasets. Gu et al. (2024) demonstrated an example of combining FinBERT with LSTM models for stock price prediction, showing that fine-tuning pre-trained models on financial data leads to more accurate predictions. This hybrid approach of using domain-specific models for sentiment analysis, combined with other learning models like LSTM, has become a standard in the field due to its ability to model both the sentiment and dependencies that drive financial markets. Another recent advancement in the financial sentiment analysis domain is the integration of other deep learning techniques such as LSTMs and CNNs with traditional sentiment lexicons. Kaliappan et al. (2023) explore this approach by analyzing financial news headlines using both sentiment lexicons and deep learning models. They found that the combination of these methods enhanced sentiment classification accuracy. This hybrid approach was found to be particularly valuable in financial sentiment analysis as headlines often contain specific terminology that general NLP models struggle with. The combination of lexicon-based methods with deep learning frames a robust method to handle the nuanced complexities of financial news. Research moving forward points to hybrid models that combine sentiment lexicons with deep learning techniques as well as hybrid models that combine multiple data sources in order to further improve financial sentiment analysis.

Deviating from the original project proposal, the development of models was deemed to be beyond the scope of reality due to the constraints of time and knowledge for the research team. The project was updated to utilize pre-built models from Huggingface instead of building the models and to continue with the plan to train models on financial news datasets. Thus, the project summarized by this paper seeks to investigate the application of BERT and LSTM style models to sentiment analysis by fine-tuning the models on datasets consisting of financial news headlines and updates. This marks a good interim learning opportunity for the author on the road to developing and implementing the most novel approaches in financial data science.

II. Method

A. Models and Data

The HuggingFace platform was used to find datasets that contain financial news and labeled sentiment and learning models to fine-tune. BERT is one of the best transformer models, but is known for being resource intensive, so a variant that is modeled after BERT but is much less resource intensive was found, DistilBERT. An LSTM model was also desirable. However, trustworthy LSTM models for sentiment analysis on Huggingface were scarce so two were chosen: a trustworthy LSTM model built on Keras and due to programming complications for metrics, another LSTM that is less reputable but compatible with Huggingface programming methods. A lexicon based sentiment analysis model was not found on

Huggingface and is not included in this project. The value and description remains in the literature review. Two datasets of financial news with labeled sentiment were found. Dataset 1 containing over 10 thousand lines was used as the main dataset and Dataset 2 was added to investigate the impact of a larger dataset on analysis. I again note that the Keras LSTM would not provide me additional metrics despite spending 2 hours debugging. I have declared it a platform issue and ingrained at a lower level I can not touch. Thus necessitating the second LSTM model.

B. Project Structure

This project compares a DistiledBERT and LSTM model as well as the impact of increasing the dataset. The questions of which model is better and if almost doubling the dataset will improve the models are investigated. The project is structured as follows:

(1) Dataset 1 is imported. Preprocessing and tokenization are applied.
(2) Initialization, training, and evaluation of the DistilBERT model.
(3) The LSTM models are initialized, trained, and evaluated.
(4) Dataset 2 is imported, cleaned, and merged with Dataset 1.
(5) The three models are trained and evaluated on this new combined dataset henceforth called Dataset 3.
(6) Comparison of the three models.

III. Model Comparison with Dataset 1

Five main metrics are investigated: loss, accuracy, precision, F1, and recall. DistilBERT has the best loss score, two thirds that of Keras LSTM and less than half of Second LSTM. DistilBERT shows the highest accuracy among the three models, significantly outperforming the Keras LSTM and Second LSTM models. The Keras LSTM model performs better than the Second LSTM model in terms of accuracy. DistilBERT has the highest F1 score, suggesting a better balance between precision and recall. The Keras LSTM model does not report an F1 score, and the Second LSTM's F1 score is significantly lower, indicating that it struggles with finding the right balance between precision and recall. DistilBERT has the highest precision, indicating that it is more accurate in identifying positive predictions. The Keras LSTM model does not provide precision data, and the Second LSTM shows very low precision, which means it misclassifies positive labels frequently. DistilBERT also performs better in recall, which indicates it is better at identifying the actual positives. The Second LSTM model shows a much lower recall, indicating it misses a significant number of positive labels. In summary, DistilBERT outperforms both the Keras LSTM and Second LSTM models across all key metrics: accuracy, F1 score, precision, and recall. Keras LSTM, only having the loss and accuracy metric, fairs decently in accuracy with 78.6%. The Second LSTM laggs far behind the DistilBERT model performance in all metrics and at times by a large margin. Only DistilBERT has crossed the 80% accuracy threshold I was looking for when planning this project.

Metric	DistilBERT	Keras LSTM	Second LSTM
Eval Loss	0.3961	0.5896 (val_loss)	0.8864
Eval Accuracy	84.6%	78.6% (val_accuracy)	65.2%
Eval F1 Score	0.7949	N/A	0.2631
Eval Precision	0.8078	N/A	0.2173
Eval Recall	0.7836	N/A	0.3333
Eval Runtime (seconds)	10.622	N/A	2.8323
Eval Samples/Second	47.074	N/A	176.538
Eval Steps/Second	3.013	N/A	11.298
Epochs	2	2	2

Fig. 1. Comparison of Sentiment Analysis Models with Dataset 1

A. Model Comparison with Datset 3

The between-model comparison remains the same when trained on this dataset. What is surprising is the by-model comparison between the training on Dataset 1 and Dataset 3. DistilBERT improved in all key metrics except in loss where it increased from 0.3961 to 0.5441. This could suggest that the model is better able to capture the complexity of the larger dataset, but the increased loss could reflect challenges in fitting to a larger and more diverse dataset. Keras LSTM did increase in its loss metric but also had a slight decrease in its accuracy. This may indicate the larger dataset may have introduced new challenges but not necessarily worse generalizations. The Second LSTM overall saw improvement in key metrics which points to this model benefiting from a larger diverse dataset. Between DistilBERT and LSTM models, improvements and maintenance were seen, encouraging the reasoning that a larger dataset could improve model performance.

Metric	DistilBERT	Keras LSTM	Second LSTM
Eval Loss	0.5441	0.7227 (val_loss)	0.7740
Eval Accuracy	87.6%	77.2% (val_accuracy)	68.8%
Eval F1 Score	0.8459	N/A	0.4349
Eval Precision	0.8340	N/A	0.4233
Eval Recall	0.8603	N/A	0.4558
Eval Runtime (seconds)	14.4446	N/A	2.5199
Eval Samples/Second	34.615	N/A	198.422
Eval Steps/Second	2.215	N/A	12.699
Epochs	2	2	2

Fig. 2. Comparison of Sentiment Analysis Models with Dataset 3

B. Review of Datasets 1 and 3

Dataset 1 is composed of almost 12 thousand rows with two columns. One column for text and another for label indicating negative, positive, or neutral sentiment. A summary table, label count table, and text column description table is displayed below. The review indicates a large majority of neutral sentiment. This suggests possible lack of sentiment examples during training samples and could decrease model ability to predict a directioned sentiment.

Dataset 3 sees an addition of 9 thousand rows of closer to half neutral and half positive or negative sentiment. This increased the diversity of labels and improved the ratio of label examples in the dataset. This can increase the models training on directioned sentiment data to better identify words or phrases that indicate sentiment. An important discovery as seen in Table 6 is the 3 thousand duplicates that Dataset 2

Metric	Value
Count	11931
Mean	1.4991
Std	0.7416
Min	0
25%	1
50%	2
75%	2
Max	2

TABLE I
SUMMARY STATISTICS OF DATASET 1

Label	Count
2	7744
1	2398
0	1789

TABLE II
LABEL DISTRIBUTION OF DATASET 1

Metric	Value
Count	11931
Unique	11931
Top	TCO, NNVC, GPOR and JE among midday movers
Frequency	1

TABLE III
TEXT COLUMN DESCRIPTION OF DATASET 1

Metric	Value
Count	20986
Mean	1.4553
Std	0.7358
Min	0
25%	1
50%	2
75%	2
Max	2

TABLE IV
SUMMARY STATISTICS OF DATASET 3

Metric	Value
Count	20986
Mean	1.4553
Std	0.7358
Min	0
25%	1
50%	2
75%	2
Max	2

TABLE V
SUMMARY STATISTICS OF DATASET 3 (TABLE V)

Label	Count
2	12632
1	5276
0	3078

TABLE VI
LABEL DISTRIBUTION OF DATASET 3 (TABLE VI)

adds. This shows clear need for data cleaning otherwise the models will be overtrained on data and fail to properly predict sentiment on the the test data. The duplicates were removed and the data refined.

IV. RESULTS

The models were evaluated on Dataset 1 and Dataset 3 to assess their performance in sentiment analysis of financial headlines. The evaluation focused on five key metrics: Accuracy, F1 Score, Precision, Recall, and Loss. The results were compared across the DistilBERT, Keras LSTM, and Second LSTM models to determine their effectiveness with varying dataset sizes.

Overall, DistilBERT was the best-performing model, achieving the highest accuracy and demonstrating the advantages of fine-tuning a pre-trained transformer model. The Second LSTM, however, showed notable improvements with the expanded dataset, highlighting its potential for further tuning and refinement. Between the two LSTM models, Keras model was found to be the better performing model. This difference can be attributed to the difference of how the authors developed the models and application of LSTM principles.

Beyond different models and expanding the dataset, other factors were not changed in testing. Future expansion of this project can investigate how fine-tuning the hyperparameters, adding in dedicated models such as FinBERT, and how temporal data such as stock prices are correlated with news in a time-series inclusion.

V. CODE REPOSITORY

The implementation details of this project are available at the GitHub Repository.

REFERENCES

[1] Gu, Wen jun, Yi hao Zhong, Shi zun Li, Chang song Wei, Li ting Dong, Zhuo yue Wang, Chao Yan, et al. "Predicting Stock Prices with Finbert-LSTM: Integrating News Sentiment Analysis." Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing. ACM, November 8, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3694860.3694870. [Accessed: Nov. 25, 2024].

[2] Kaliappan, S., L. Natrayan, and Akshay Rajput. "Sentiment Analysis of News Headlines Based on Sentiment Lexicon and Deep Learning." Proceedings of the 2023 IEEE International Conference on Smart Cities and Emerging Technologies (ICOSEC), 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10276102. [Accessed: Nov. 25, 2024].

[3] Hugging Face, "zeroshot/twitter-financial-news-sentiment Dataset," Hugging Face, [Online]. Available: https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment. [Accessed: Nov. 25, 2024].

[4] Hugging Face, "distilbert-base-uncased Model," Hugging Face, [Online]. Available: https://huggingface.co/distilbert/distilbert-base-uncased. [Accessed: Nov. 25, 2024].

[5] Hugging Face, "RobertoMCA97/financial_sentiment_analysis_train_compilation Dataset," Hugging Face, [Online]. Available: https://huggingface.co/datasets/RobertoMCA97/financial_sentiment_analysis_train_compilation. [Accessed: Nov. 25, 2024].

[6] Hugging Face, "keras-io/bidirectional-lstm-imdb Model," Hugging Face, [Online]. Available: https://huggingface.co/keras-io/bidirectional-lstm-imdb. [Accessed: Nov. 25, 2024].

[7] Hugging Face, "LYTinn/lstm-finetuning-sentiment-model-3000-samples Model," Hugging Face, [Online]. Available: https://huggingface.co/LYTinn/lstm-finetuning-sentiment-model-3000-samples. [Accessed: Nov. 25, 2024].