Stock Sentiment Prediction Using Financial News Data

R. Permana, *Student Member*, and A. I. Lukito, *Student Member*, IEEE
School of Computer Science, Bina Nusantara University, Indonesia

***Abstract*—This research examines the effectiveness of deep learning techniques for automated sentiment prediction in the stock market, utilizing financial news articles, and addresses the inefficiency of manual analysis due to the volume of daily articles. To achieve this, financial news articles are collected and preprocessed to train and compare supervised deep learning architectures, specifically transformer-based models (such as FinBERT, ALBERT) and long short-term memory (LSTM) models combined with word embedding (such as Word2Vec, FastText), to predict continuous sentiment scores from the news content. Model performance was rigorously evaluated using standard regression metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and the R^2 score, with experimental findings demonstrating that transformer-based models, like FinBERT, exhibit superior performance over traditional sequential models in accurately capturing the contextual and semantic nuances present in financial text.**

***Index Terms*—Sentiment Analysis, Stock Market, Deep learning, Transformers, LSTM, Word Embedding, Finance**

I. INTRODUCTION

With the advancement of modern technology, information has become more accessible than ever. Over the years, news has played a significant role in influencing stock market volatility and uncertainty. Investors often rely on news coverage when making decisions about which stocks to buy, as human behavior tends to be guided by information that captures their attention. (Xu, 2025).

Given the vast amount of news that must be analyzed, a sentiment analysis model can help determine whether a stock is viewed positively or negatively by the public. A supervised deep learning approach is well-suited for this task, as previous research has demonstrated that transformer-based models, such as BERT and its derivatives, are effective in understanding and capturing semantic meaning and contextual relationships within text (Li & Hu, 2025, p. 13).

This study aims to identify and evaluate the most effective model for detecting and quantifying the sentiment of stocks mentioned in financial news. By accurately assessing sentiment from news-based data, this research contributes to the development of more efficient financial analytics tools that can enhance decision-making and market insight.

II. LITERATURE REVIEW

A. Deep Learning

Deep learning is a subset of machine learning that works through a multi-layer neural network, inspired by the human brain (Bergmann, n.d.). Each layer in the neural network contains several neurons that perform mathematical calculations using input values, weights, and biases to yield an output (Bergmann, n.d.). The output would be evaluated and used to adjust the weights and biases through backpropagation, thereby improving the quality of the prediction.

B. Text Mining

Text mining is a process of converting unstructured text data into a structured format to uncover patterns, hidden links, and key ideas (IBM, n.d.). In this research, utilizing text mining would be beneficial in analyzing vast amounts of news to extract sentiment that may influence stock market behavior.

C. Web Scraping

Web scraping is collecting publicly accessible content from a website and storing the data in a database, file, or spreadsheet for later analysis (Brave, 2024). It utilizes bots (or Web Crawlers). In which the bot and crawlers are designed to visit multiple websites or pages within a website and collect data about the contents of those sites (Brave, 2024).

D. Stock Market

The Stock market is a broad term for a network of exchanges where investors buy and sell publicly traded shares (Gratton, 2025). With this in mind, it is beneficial to utilize financial news to identify how stock market–related information influences investor sentiment, stock price movements, and overall market volatility.

E. Sentiment Analysis

As there are numerous news stories to analyze, sentiment analysis is beneficial for producing a concise summary of the sentiment score related to a stock. Sentiment analysis, also known as opinion mining, utilizes natural language processing to automatically extract and analyze sentiment from a given text (Chaturvedi et al., 2018; Liu & Zhang, 2012, as cited in Mao et al., 2024).

F. BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a model that is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without requiring substantial task-specific architecture modifications (Devlin et al., 2018, p. 1).

G. BERTopic

BERTopic is a topic modeling method proposed by Grootendorst in 2022, which has demonstrated effective topic identification in various domains (Egger & Yu, 2022; Grootendorst, 2022; Jeon et al., 2023, as cited in Li & Hu, 2025). BERTopic captures semantic relationships more accurately and adapts well to unstructured text with complex contextual patterns. Its flexibility and high-quality topic outputs make it widely used in research involving text classification, trend analysis, and thematic exploration.

H. FinBERT

FinBERT is a subset of BERT that is trained on a large financial corpus and fine-tuned for sentiment classification (Hugging Face, n.d.). Research indicates that FinBERT outperforms conventional deep learning models, such as LSTM, in classifying financial data (Araci, 2019, p. 6; Gu et al., 2024, p. 8). Due to this domain specialization, FinBERT has become a strong baseline model in financial NLP tasks, serving as a benchmark for evaluating other sentiment analysis approaches in this study.

I. ALBERT

As its name suggests, A Light Bert (ALBERT) is a pre-trained model developed to reduce training time and GPU/CPU memory usage while maintaining high performance (Lan et al., 2020, pp. 1-2). Such high performance with low training time is achieved by implementing factorized embedding parameterization and cross-layer parameter sharing (Lan et al., 2020, 2).

J. LSTM

Long Short-Term Memory (LSTM) networks were first introduced by Hochreiter and Schmidhuber in 1997 and later refined by numerous researchers. LSTM has become one of the most widely used architectures for sentiment classification due to its ability to capture long-term dependencies in sequential data (Murthy et al., 2020, p. 299). In this project, the LSTM model will be further improved through hyperparameter tuning using Optuna, a next-generation optimization framework developed by Takuya Akiba in 2019, to generate sentiment scores for evaluation.

K. Hyperparameter Tuning

Hyperparameter tuning is a critical aspect of machine learning that greatly influences model performance, as proper tuning—ranging from simple grid search to advanced Bayesian optimization and meta-learning—can substantially improve accuracy, efficiency, and generalization while balancing computational cost and performance gains (Ilemobayo et al., 2024, p. 393).

As stated previously, Optuna will be utilized for hyperparameter tuning of the LSTM model. It is chosen due to its capabilities of efficient implementation in searching and pruning strategies (Akiba et al., 2019, p. 1). Through iterative trial evaluations, Optuna finds the optimal combination of parameters–such as learning rate, number of units, dropout rate, and batch size–that improves model performance.

L. Word2Vec

Word2Vec is a word embedding model developed by Google, which aims to map words to high-dimensional vectors to capture the semantic relationships between words. It utilizes two main architectures: Continuous Bag-of-Words (CBOW), which predicts a target word from surrounding context words, and Skip-gram, which predicts context words from a target word (GeeksforGeeks, 2025).

M. FastText

FastText, created by Facebook's AI Research (FAIR) lab, extends Word2Vec by incorporating subword information through character n-grams (typically 3-6 characters). This allows it to generate embeddings for rare, misspelled, or out-of-vocabulary words by summing vectors of their n-grams, such as breaking "apple" into "app", "ppl", "ple". It supports the same Skip-gram and CBOW modes, similar to Word2Vec, and excels in handling noise better than word-level models (GeeksforGeeks, 2025).

A. Topic

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras vel purus vel dui molestie bibendum vel venenatis sem. Nam condimentum ipsum ex, suscipit imperdiet elit pretium in. Fusce feugiat semper rutrum. Aenean elementum ipsum hendrerit nisl tempor, gravida blandit felis facilisis. Aenean in pretium ligula. Proin mollis sed mi a ornare. Praesent nec quam volutpat, dictum erat sit amet, pulvinar elit. Nam magna orci, mattis quis leo aliquet, dictum volutpat massa. Nulla auctor, ex at imperdiet viverra, eros est porta purus, sed ullamcorper neque risus et ligula. Nunc dolor enim, porta nec quam vitae, ultrices mattis enim. Vestibulum bibendum sem ac dapibus aliquet. Aliquam nisl arcu, sodales at tellus ac, malesuada dapibus dui. Nulla venenatis lacus quis quam dapibus dignissim.

III. METHODOLOGY

3.1 Data Collection

Financial news articles were collected from Marketaux API. sources, including reputable financial media outlets and news aggregators, with metadata, including the publication date, headline, and full text, accompanying each article. To link the news to specific companies, the dataset was filtered using company names and stock ticker symbols. Ensuring that

each article could be associated with one or more relevant stocks. Duplicate articles and irrelevant news were removed through a combination of keyword filtering and manual inspection.

3.2 Data Preprocessing

Before model training, all textual data undergo several preprocessing steps. The text was first converted to lowercase and cleaned by removing punctuation, URLs, stopwords, and non-alphabetic characters. Tokenization was performed using a transformer-compatible tokenizer, preserving the semantic structure of the text. Lemmatization was also applied to standardize word forms.

3.3 Model Architecture

The core model used in this study was a transformer-based language model specifically optimized for understanding financial text. Variants of BERT and its derivatives (such as FinBERT and ALBERT) were evaluated for their ability to capture contextual and semantic relationships between words. Simpler models, such as the Long Short-Term Memory (LSTM) model, will also be compared. The model input consisted of processed news headlines, article summaries, and news article content, and the output was a continuous sentiment score ranging from -1 (negative one) to 1 (one), representing the predicted market sentiment for each stock mentioned in the text.

3.4 Training and Optimization

The dataset was divided into training (70%), validation (10%), and testing (20%) subsets. The model was fine-tuned using a supervised approach with mean squared error (MSE) as the loss function, since the target output was a continuous sentiment score. AdamW optimizer was used with a learning rate scheduler to ensure stable convergence. Early stopping was applied to prevent overfitting. Hyperparameters, such as learning rate, batch size, and maximum sequence length, were optimized through a grid search.

3.5 Evaluation

Model performance was evaluated using multiple regression and classification metrics. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to measure the accuracy of sentiment score predictions. For additional interpretability, the predicted sentiment scores were classified into positive, neutral, and negative categories using predefined thresholds and evaluated using Precision, Recall, and F1-score.

3.6 Implementation

All experiments were implemented in Python using the PyTorch framework. Transformers were utilized through the Hugging Face library, and text preprocessing was conducted using NLTK and regular expressions. Data handling and visualization were performed using the Pandas and Matplotlib libraries. The model was trained on Google Colab using an Nvidia GPU to accelerate training and evaluation.

IV. Results And Discussion

Model	MSE	MAE	RMSE	r^2
FastText + LSTM	0.0895	0.245	0.2991	0.0042
Word2Vec + LSTM	0.0932	0.2539	0.3054	-0.0043
Albert-v2	0.0888	0.2182	0.298	0.0917
FinBert	0.0844	0.2066	0.2905	0.1093

Table 1 Model Score Results

After performing hyper-parameter tuning and training, the performance of the evaluated models was assessed using regression metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), as shown in Table 1. Among all tested approaches, FinBERT achieved the strongest overall performance, yielding the lowest MSE, MAE, and RMSE values, as well as the highest R^2 score.

These results indicate that FinBERT produces more accurate sentiment score predictions compared to the other evaluated models. Most likely due to the model being trained on a financial corpus and fine-tuned for sentiment classification (Hugging Face, n.d.). However, despite its relative performance advantage, the R^2 value remains low, suggesting that a substantial portion of sentiment variability in financial news data is not captured by the model. This reflects the inherent complexity and noise present in financial text and market-related sentiment signals.

Overall, FinBERT provides the most reliable sentiment prediction among the compared models in this study. The results suggest that while the model is effective for capturing general sentiment patterns in financial news, its predictive capability is more suitable for aggregate sentiment analysis rather than precise, article-level sentiment prediction.

V. Conclusion

This study evaluated several deep learning models and transformers for predicting stock market sentiment from financial news articles. The results show that FinBERT outperforms the other evaluated models, achieving the lowest error metrics and the highest R^2 value, which indicates its effectiveness in capturing aggregate sentiment patterns in news financial text. However, the relatively low R^2 scores across all models suggest that more data is needed, an increase in training iterations, or financial data alone is insufficient.

Despite these limitations, the proposed approach provides a scalable framework for extracting stock market sentiment from large volumes of news data. Future work may focus on improving sentiment prediction through additional features such as Named Entity Recognition, and combining with time series data, which may enhance the analysis of

sentiment–price relationships and improve the practical usefulness of sentiment-based market analytics.

References

Akiba, T., Sano, S., Yanase, T., Otha, T., & Koyama, M. (2019, July 26). Preprint doi Optuna: A Next-generation Hyperparameter Optimization Framework. https://arxiv.org/abs/1907.10902

Araci, D. T. (2019, 06 05). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. https://arxiv.org/pdf/1908.10063

Bergmann, D. (n.d.). *What is deep learning?* What is deep learning? | IBM. Retrieved October 15, 2025, from https://www.ibm.com/think/topics/deep-learning

Brave. (2024, July 10). *Web Scraping Meaning & Definition*. Brave. Retrieved October 20, 2025, from https://brave.com/glossary/web-scraping/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

GeeksforGeeks. (2025, July 23). *Word Embeddings Using FastText*. GeeksforGeeks. Retrieved December 19, 2025, from https://www.geeksforgeeks.org/nlp/word-embeddings-using-fasttext/

GeeksforGeeks. (2025, October 4). *Word Embedding using Word2Vec*. GeeksforGeeks. Retrieved December 19, 2025, from https://www.geeksforgeeks.org/python/python-word-embedding-using-word2vec/

Gratton, P. (2025, July 15). *What Is the Stock Market and How Does It Work?* Investopedia. Retrieved October 20, 2025, from https://www.investopedia.com/terms/s/stockmarket.asp

Gu, W., Zhong, Y., Li, S., Wei, C., Dong, L., Wang, D., & Yan, C. (2024, 08). Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis. 67–72. http://dx.doi.org/10.1145/3694860.3694870

Hugging Face. (n.d.). *ProsusAI/finbert · Hugging Face*. Hugging Face. Retrieved November 17, 2025, from https://huggingface.co/ProsusAI/finbert

IBM. (n.d.). *What Is Text Mining?* IBM. Retrieved October 15, 2025, from https://www.ibm.com/think/topics/text-mining

Ilemobayo, J. A., Durodola, O., Alade, O., Awotunde, O. J., Olanrewaju, A. T., Falana, O., Ogingbire, A., Osinuga, A., Ogunbiyi, D., Ifeanyi, A., Odezuligbo, I. E., & Edu, O. E. (2024, 6 7). Hyperparameter Tuning in Machine Learning: A Comprehensive Review. *Journal of Engineering Research and Reports*, *26*(6), 388-395.

Lan, Z. Z., Chen, M. D., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020, February 09). ALBERT: A Lite Bert For Self-Supervised Learning of Language Representations. *ICLR 2020*, 1-17. https://doi.org/10.48550/arXiv.1909.11942

Li, C., & Hu, X. (2025, May 27). Medical Artificial Intelligence in Scholarly and Public Perspective:

BERTopic-Based Analysis of Topic-Sentiment Collaborative Mining. *Data Science and Informetrics*. https://doi.org/10.1016/j.dsim.2025.05.001

Mao, Y., Liu, Q., & Zhang, Y. (2024, April 4). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, *36*(4). https://doi.org/10.1016/j.jksuci.2024.102048

Murthy, D. G. S. N., Allu, S. R., Andhavarapu, B., Bagadi, M., & Belusonti, M. (2020, May). Text-based Sentiment Analysis using LSTM. *International Journal of Engineering Research & Technology (IJERT)*, *09*(05). https://www.researchgate.net/publication/341873850_Text_based_Sentiment_Analysis_using_LSTM

Xu, F. (2025). *Impact of news coverage on the financial market*. Impact of news coverage on the financial market. Retrieved 10 15, 2025, from https://www.brunel.ac.uk/research/projects/impact-of-news-coverage-on-the-financial-market