Prediction of Stock Market Movements Using Natural Language Processing Techniques on Financial News and Market Events Data

Volkan İnce[#1], Yusuf Batur[#2]

[#]*Department of Engineering, Artificial Intelligence, Gebze Technical University*
Cumhuriyet, 2254. Sk. No:2, 41400 Gebze/Kocaeli, Türkiye

[1]v.ince2021@gtu.edu.tr

[2]m.batur2025@gtu.edu.tr

Abstract— **The project is designed to analyze and measure the impact of financial news on stock market movements by applying advanced Natural Language Processing (NLP) techniques to the "Financial News and Market Events Dataset 2025." The main goal is to develop a robust model that can extract measurable sentiment, identify key topics affecting market movements, and recognize named entities from unstructured news texts. By systematically correlating this extracted textual information with historical stock market data, the project will investigate the predictive power of media sentiment on asset price fluctuations. The proposed methodology includes a data preprocessing pipeline, sentiment analysis using domain-specific pre-trained language models such as FinBERT, and topic modeling using Latent Dirichlet Allocation (LDA). The main goal is to demonstrate that NLP-focused analysis can serve as a valuable and powerful complementary tool to traditional financial forecasting methods, providing investors, analysts, and risk managers with timely, data-driven, and actionable insights.**

Keywords— **Sentiment Analysis, Financial Analysis, Stock Market Prediction, Machine Learning, Topic Modeling, FinBERT, Algorithmic Trading.**

I. INTRODUCTION

A. Main Objective

The primary objective of the project is to design, implement, and evaluate a framework for analyzing high-volume financial news using Natural Language Processing (NLP) and investigating its causal relationship with stock market movements. Leveraging Kaggle's "Financial News and Market Events Dataset 2025," the project aims to convert qualitative text data (e.g., "the company reported stronger-than-expected earnings") into quantitative metrics (e.g., a +0.85 positive sentiment score). These metrics are then statistically modeled with quantitative market data (e.g., a 2% increase in stock price) to identify predictive patterns. The final output will be an analytical framework capable of providing insights into potential market trends based on real-time news.

B. Motivation and Rationale

Financial markets are highly sensitive, complex adaptive systems that respond to the flow of new information. A corporate merger, a regulatory announcement, or a change in macroeconomic sentiment (or a tweet) can trigger significant investor reactions and market volatility within minutes. While traditional financial analyses (technical analysis of price charts and fundamental analysis of company finances) provide solid forecasting, these traditional analyses often struggle to systematically incorporate the vast and unstructured data found in news articles, press releases, and social media [1]. This gap presents us with a significant opportunity.

The motivation for our project stems from the general problem of information overload in the digital age. Tens of thousands of financial news articles are published every day, making manual analysis impossible. This project aims to fill this gap by automatically processing and interpreting this massive text corpus using NLP and machine learning models. This will allow us to test one of the fundamental principles of behavioral finance: market sentiment reflected in the media is a leading indicator of market direction; this concept has previously been proven using social media data [2]. This study aims to go beyond the evidence and provide a data-driven answer to the question, "Can the collective sentiment of financial news predict stock market direction?"

II. METHODOLOGY

This section outlines the comprehensive methodological framework designed to investigate the predictive power of financial news on stock market index movements. The methodology is structured as a multi-stage process, beginning with data collection and feature engineering, followed by the application of distinct machine learning architectures for classification and regression tasks, and concluding with a robust quantitative evaluation protocol.

A. Data Corpus and Feature Engineering Framework

The foundation of any predictive modeling endeavor is the quality and structure of the input data. This subsection details the primary dataset, its preparatory procedures, and the sophisticated feature engineering pipeline designed to transform unstructured textual data and quantitative metrics into a format suitable for machine learning algorithms.

1) Dataset Characterization and Temporal Alignment: This study will exclusively use the "Financial News Market Events Dataset 2025" available on the Kaggle platform [3]. The synthetic dataset contains 3,024 records spanning from February 1, 2025, to August 14, 2025. The core columns of the dataset include Date, Headline, Market_Index, Index_Change_Percent, and Trading_Volume. The dataset's description notes that it strategically includes approximately 5% null values to simulate real-world data cleaning scenarios. As a first step, an imputation strategy will be applied to these missing values. For missing values in the numerical Index_Change_Percent and Trading_Volume columns, median imputation will be used to mitigate the impact of potential outliers. Null entries in the categorical Headline column will be removed from the dataset as they contain no predictive information. The Index_Change_Percent column will serve as the primary target variable for both prediction tasks: For the classification task (Directional Prediction): A binary target variable named Movement_Direction will be created. A value of 1 (Up) will be assigned if Index_Change_Percent > 0, and 0 (Down) if Index_Change_Percent <= 0.

For the regression task (Magnitude Prediction): The raw Index_Change_Percent value will be used directly as the continuous target variable.

A critical step in the methodology is ensuring a strict temporal relationship between the features and the target variable. Features from a given day t (news headline, trading volume) will be used to predict the market outcome on the same day t. The dataset's structure, which associates a daily headline with that day's index change, naturally facilitates this alignment.

2) Natural Language Processing for Sentiment and Semantic Feature Extraction: This stage is central to the research, focusing on converting the unstructured Headline text into meaningful quantitative features. A literature-based comparative approach will be adopted to select the most suitable NLP model. Prior to feature extraction, all news headlines will undergo a standard preprocessing sequence: convert text to lowercase, remove punctuation and special characters, tokenize the text into words and remove common English stop words. To ensure semantic consistency, lemmatization will be applied to reduce words to their root forms. These preprocessing steps, particularly stop-word removal and lemmatization, are standard practices to improve the accuracy and efficiency of NLP tasks by reducing noise and normalizing words to their base forms [4]. The literature reveals a clear distinction between lexicon-based models and modern transformer-based architectures for financial sentiment analysis. Rule-based models like VADER (Valence Aware Dictionary for sEntiment Reasoning) use a predefined dictionary of words with sentiment scores. While computationally efficient, these models often fail to capture the nuanced and context-dependent nature of financial language, where terms like "liability" or "volatility" can carry non-negative meanings [6]. In contrast, models like FinBERT are BERT models pre-trained on a large corpus of financial texts (e.g., news articles, corporate reports). This domain-specific training allows for a deep understanding of context. Studies consistently show that FinBERT significantly outperforms lexicon-based methods like VADER and general-purpose models in financial sentiment classification tasks, achieving higher accuracy and F1-scores [5].

Based on this overwhelming evidence in the literature, the FinBERT model will be used in this study. The primary justification for this choice is its ability to overcome the semantic shortcomings of lexicon-based models by interpreting words within their specific financial context, thereby producing more accurate and reliable sentiment features [6]. Two distinct feature sets will be extracted from the FinBERT model for each headline:

Sentiment Scores: A categorical feature representing the predicted sentiment (positive, negative, or neutral) and a continuous composite score.

Contextual Embeddings: A high-dimensional numerical vector will be generated for each headline. This embedding represents the semantic meaning of the headline in a dense vector space, capturing far more nuance than a single sentiment score..

3) Construction of the Hybrid Input Vector: Academic research has shown that the performance of prediction models significantly improves when textual features are combined with quantitative market data [7]. Therefore, a hybrid feature vector will be created for each data instance. The final input vector for the models will combine features from multiple sources: the sentiment score derived from FinBERT, the Trading_Volume provided in the dataset, a lagged variable named Previous_Day_Change representing the Index_Change_Percent from the previous day (*t-1*) to capture momentum effects, and the 768-dimensional contextual embedding from FinBERT, which will serve as the primary input to capture deep textual meaning.

The primary dataset used in this study , while rich in news-outcome linkages, lacks the traditional OHLC (Open, High, Low, Close) price data commonly found in other stock market datasets [3]. This prevents the calculation of standard technical indicators such as RSI or MACD, which are frequently used as features in the stock prediction literature [9]. This constraint necessitates a methodological pivot. Instead of relying on a broad set of technical indicators, the model's predictive power will heavily depend on the quality and richness of the features extracted from the news headlines. Therefore, the choice to

use FinBERT to generate high-dimensional contextual embeddings, rather than just a simple sentiment score, becomes critically important. These embeddings must act as a proxy for the market-moving information that technical indicators would otherwise capture. This study is thus designed to indirectly test the hypothesis of whether deep semantic features from news text can compensate for the absence of traditional technical indicators in financial forecasting.

B. Predictive Modeling Architectures

Next step is to select and design the machine learning models. This study proposes a dual-task approach, employing a robust ensemble model for the classification task and a comparative analysis between two different architectures for the more challenging regression task.

1) *Directional Market Prediction: An Ensemble Classification Approach:* For the binary classification task of Movement_Direction (up/down), an Extreme Gradient Boosting (XGBoost) classifier will be implemented. XGBoost is a tree-based ensemble learning algorithm known for its high performance, speed, and scalability. It has shown strong results in stock market prediction, effectively handling tabular data with heterogeneous feature types (numerical, categorical) and demonstrating resilience to overfitting [10]. Research has shown that XGBoost can achieve high accuracy (e.g., 73% in a similar task) even with only price-derived features, suggesting its potential will be further enhanced by our rich, NLP-driven feature set [11].

2) *Magnitude of Change Prediction: A Comparative Regression Analysis:* Predicting the exact Index_Change_Percent requires a more nuanced approach. This study will implement and compare two powerful yet fundamentally different regression models to determine which architectural assumption is better suited to the problem.

Model A: XGBoost Regressor: As a direct counterpart to the classification model, an XGBoost regressor will be trained. This model will treat each day's feature vector as a static, independent input to predict the corresponding continuous outcome. Its inclusion is justified by its documented success in regression tasks and its high performance in various data science competitions [12].

Model B: Long Short-Term Memory (LSTM) Network: Acknowledging the inherently time-series nature of financial markets, an LSTM model will be implemented as a comparative approach. LSTMs are a type of Recurrent Neural Network (RNN) specifically designed to learn long-term dependencies in sequential data [13]. The LSTM will be trained on sequences of hybrid feature vectors from the past n days (e.g., a 5-day sliding window) to predict the next day's outcome. This architecture is specifically designed to capture temporal patterns and momentum that a static model like XGBoost cannot [7]. The model will use a hybrid input structure, similar to that described in the literature, where numerical and textual features are combined at each time step [8].

The choice between XGBoost and LSTM for the regression task represents more than just a model comparison; it is a test of a fundamental hypothesis about market dynamics. The XGBoost model operates under the assumption that the most relevant information for tomorrow's market movement is contained within today's complete feature set (today's news, volume, yesterday's close, etc.). It models a static, instantaneous relationship. The LSTM model, in contrast, operates under the assumption that the sequence and evolution of features over the last few days hold critical predictive power. It assumes, for example, that three days of increasingly negative headlines are more informative than a single day's news. By implementing both, we are not only comparing algorithms but empirically testing whether stock market magnitude is better modeled as a static function of current information or as a dynamic function of recent temporal patterns. The results of this comparison will offer a deeper insight into the nature of financial data.

3) *Model Implementation, Training, and Optimization:* To prevent data leakage and simulate a realistic forecasting scenario, the dataset will be split chronologically. The first 70% of the data will be used for training, the next 15% for validation and hyperparameter tuning, and the final 15% will be held out as a blind test set.

The optimal hyperparameters for both the XGBoost and LSTM models are data-dependent. A systematic search will be conducted using a Bayesian Optimization approach on the validation set. This method is more efficient than exhaustive Grid Search or Random Search at finding optimal model configurations. Key parameters to be tuned will include the learning rate, tree depth (for XGBoost), and the number of hidden units and layers (for LSTM).

C. Analysis, Correlation, and Modeling

1) Time Series Construction: The outputs from the NLP models will be aggregated to create daily time series data, such as a "Market Sentiment Score" (average sentiment of all news) or a "Tesla-Specific Sentiment Score" (average sentiment of news mentioning TSLA).

*2) Statistical Analysis***:** Correlation analysis will be performed to measure the statistical relationship between the sentiment time series and the stock price time series (e.g., daily returns). Furthermore, a Granger Causality Test will be conducted to investigate whether the sentiment score time series is a statistically useful predictor of the stock price time series, implying a potential predictive relationship.

3) Feature Engineering and Predictive Modeling (Optional Extension): If time permits, the sentiment scores, topic distributions, and other extracted features will be used to train a simple machine learning model (e.g., Logistic Regression or a Long Short-Term Memory network - LSTM) to predict the next day's market direction (up or down).

III. Expected Results

1) A quantifiable, statistically significant lead-lag relationship between spikes in positive/negative news sentiment and subsequent stock price movements within a defined time window (e.g., 24-48 hours).

2) Identification and visualization of the dominant market themes for different periods covered by the dataset, and an analysis of how these themes correlate with sector-specific market performance.

3) A robust, documented Python codebase (likely in a Jupyter Notebook) that performs the entire analysis pipeline from data ingestion to final visualization.

4) Empirical evidence demonstrating that news related to specific event types (e.g., M&A announcements, earnings reports, regulatory changes) has a measurably higher and more immediate impact on volatility and price than general market commentary.

IV. Possible Benefits

*1) Academic Contribution***:** The project will contribute to the field of computational finance and the ongoing debate on the Efficient Market Hypothesis by providing a modern, NLP-based case study on the role of information in price discovery.

*2) Practical Application & Commercial Value***:** The developed framework can serve as a prototype for sophisticated financial tools. Hedge funds, asset managers, and retail investors could use such technology for alpha generation (identifying trading opportunities) and risk management (detecting rising negative sentiment around a specific asset).

3) Technical Skill Development: The project provides hands-on experience in highly sought-after skills, including applied machine learning, big data processing, time series analysis, and working with state-of-the-art NLP models.

References

[1] T. H. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35-65, 2011.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1-8, 2011.

[3] P. Puri. (2025). *Financial news market events dataset for NLP 2025*. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/pratyushpuri/financial-news-market-events-dataset-2025

[4] N. Al-Madi, M. Al-Zewairi, I. Almomani, and M. A. Al-Betar, "A novel hybrid method for stopwords identification," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 9, pp. 1125–1135, 2021.

[5] L. Zhang, S. Wang, and B. Wang, "A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting," *Mathematics*, vol. 10, no. 3, p. 370, 2022.

[6] F. Saeed, D. Blyth, and Y. Guo, "An analysis of different sentiment analysis models on financial text using transformer," in *Proc. 2023 6th International Conference on Information and Computer Technologies*, 2023, pp. 148-154.

[7] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, "Deep learning for stock prediction using numerical and textual information," in *Proc. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016, pp. 1–6.

[8] W. Khan, F. Ali, A. Khattak, M. Jibran, and M. Shah, "LSTM based stock prediction using weighted and categorized financial news," *PLOS ONE*, vol. 18, no. 3, p. e0282234, Mar. 2023.

[9] G. Mneirji and F. Göransson Hörnfeldt. (2025). *A machine learning-based stock prediction system using XGBoost*. [Online]. Available: https://kth.diva-portal.org/smash/get/diva2:1985833/FULLTEXT01.pdf

[10] Z. Li, J. Li, and J. Li, "Stock index direction forecasting using an explainable eXtreme Gradient Boosting and investor sentiments," *The North American Journal of Economics and Finance*, vol. 64, p. 101848, 2022.

[11] I. E. Livieris, N. Kiriakidou, and S. Stavroyiannis. (2021). *Predicting the movement direction of OMXS30 stock index using XGBoost and sentiment analysis*. [Online]. Available: http://www.diva-portal.org/smash/get/diva2:1531990/FULLTEXT02.pdf

[12] A. Gifty and Y. Yang, "A comparative analysis of LSTM, ARIMA, XGBoost algorithms in predicting stock price direction," *ETJ*, vol. 9, no. 08, pp. 4978-4991, 2024.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.