Market Sentiment Analysis in Financial News and Stock Price Prediction

Introduction to Information Retrieval and Text Mining 2024 Fall final project

Siang Ruei Hu
dept. Information Management
National Taiwan University
b11705028@ntu.edu.tw

Yi Ting Chen
dept. Information Management
National Taiwan University
b11705051@ntu.edu.tw

Chen Hsiang Liao
dept. Information Management
National Taiwan University
b11705058@ntu.edu.tw

Yen Hung Chiang
dept. Information Management
National Taiwan University
b11705044@ntu.edu.tw

Ting Yu Chen
dept. Information Management
National Taiwan University
b11705053@ntu.edu.tw

***Abstract*—With the advancement of information technology, financial news has become more and more important in investment. We want to use some method to analyze financial news and try to predict stock price fluctuations. By leveraging natural language processing techniques, we can capture the sentiments within new content, quantifying these insights to support a stock price prediction model.**

***Index Terms*—Stock Price Prediction, Market Sentiment Analysis, Text Mining, Machine learning**

I. Introduction

The financial market is a dynamic and complex environment where decisions are strongly influenced by the vast flow of information circulating daily. Within these, financial news plays a crucial role in shaping investor sentiment and driving market movements. We want to know whether the sentiment expressed in financial news affects stock market trends. Thus, with the rapid progress of information technology, there is now an unprecedented opportunity to systematically analyze news articles, uncovering patterns and insights that can help predict stock price fluctuations. To achieve this, we employ various models such as Random Forest, XGBoost, LightGBM, CatBoost, and others to examine whether today's positive or negative news sentiment correlates with tomorrow's stock price trends. Additionally, we analyze the results and find what factors may lead to the consequences.

II. Related Work

Research on market sentiment analysis and its relationship to stock price prediction has been a significant focus for years. Early studies primarily used lexicon-based approaches, utilizing predefined sentiment dictionaries specifically tailored for financial contexts. Over time, researchers also began to train models such as logistic regression and support vector machines on labeled datasets to classify sentiment in financial texts, achieving moderate success. The influence of sentiment on stock prices has been extensively studied, with findings indicating that positive or negative sentiment often correlates with short-term price movements. Therefore, we want to think that different models will exhibit multiple performance due to the inherent differences in their underlying algorithms.

III. Approach

Our approach to market sentiment analysis in relation to stock prices utilizes FinBERT, a financial domain-specific adaptation of the BERT language model.

A. FinBERT

The key advantage of FinBERT lies in its specialization in analyzing financial texts. Unlike general-purpose models, FinBERT has been pre-trained on a vast corpus of financial documents, making it particularly adept at understanding domain-specific terms such as "interest rates," "economic events," and "stock trends," along with their implications for market sentiment. When performing sentiment analysis, FinBERT processes headlines and content in batches, generating precise sentiment scores for both positive and negative statements. These scores provide valuable insights into how financial news and events might influence the market. By leveraging FinBERT, we can efficiently analyze large volumes of financial data with higher accuracy, making it a powerful tool for predicting stock prices, assessing risks, and informing decision-making in the financial sector.

B. Main Approach

After processing the data with FinBERT, we then start doing training with the following models.

For stock price prediction, we employed six distinct models:

1. Random Forest
2. Multi-Layer Perceptron (MLP)
3. XGBoost
4. LightGBM (LGBM)
5. CatBoost
6. Stacking

1) Model 1: Random Forest: Random Forest (RF) is an ensemble method based on decision trees. It uses multiple trees to perform feature selection and classification. RF does not make additional transformations or assumptions about the input data, making it capable of directly accepting the numerical features output by FinBERT. Its advantage lies in its ability to handle nonlinearly separable datasets.

2) Model 2: MLP: Model 2 uses the Multi-Layer Perceptron (MLP) model for training. MLP is a neural network-based model that is well-suited for handling high-dimensional data. The embeddings provided by FinBERT can be directly used as the input layer of MLP. Through multiple fully connected layers, the model learns features and performs classification. MLP is particularly effective in processing complex feature patterns and high-dimensional data. It can include nonlinear activation functions to learn higher-level features. However, its disadvantages include slower training speed and a higher tendency to overfit.

3) Model 3: XGBoost: Model 3 uses the XGBoost model for training. XGBoost is a gradient-boosting decision tree model that learns the relationship between input features and target variables by utilizing tree structures. By incrementally adding trees (one at a time), XGBoost effectively captures complex interactions between features. Its advantages include efficient computation and memory utilization, as well as the ability to handle nonlinearly separable data and complex feature interactions.

4) Model 4: LGBM: Model 4 utilizes LightGBM (LGBM). LGBM is a fast implementation of gradient-boosting decision trees. It uses a leaf-wise growth strategy to construct decision trees, enhancing training speed and memory efficiency. It is particularly suitable for high-dimensional and large-scale data. The advantages of LGBM include faster training speed compared to XGBoost and support for handling sparse features (e.g., missing values). However, it may perform inconsistently on small datasets, requiring appropriate parameter tuning.

5) Model 5: CatBoost: The fifth model employs CatBoost. CatBoost specializes in handling categorical features and missing values, making it particularly effective for datasets with categorical data. When processing the continuous numerical features output by FinBERT, CatBoost operates similarly to XGBoost and LGBM by gradually building an ensemble of boosted trees. Its advantages include automatic handling of categorical features and reduced risk of overfitting, leading to stable performance on small datasets. However, its training speed is slightly slower compared to LGBM.

6) Model 6: Stacking: The final method is Stacking. This approach combines the predictions of multiple models (e.g., RF, XGB, LGBM, CatBoost) as new features. These predictions are then input into a meta-model, which typically performs the final prediction. Stacking can leverage the strengths of multiple models to improve generalization performance. By utilizing the diversity of base models, it

reduces the bias and variance of a single model. However, its drawbacks include a more complex training process and higher computational cost. If the base models perform poorly, the overall performance of the stack may be adversely affected.

By integrating these six models, our main approach aims to leverage the strengths of each model to build a system that predicts stock prices based on sentiment scores.

IV. Implementation

A. Data

About data collection, we categorized the data into two types: the first is news sources, and the second is stock price trends. Next, we will explain how we obtained these two types of data.

1) **News Source:** The dataset we collected comes from a variety of sources to make sure the information is well-covered. News data was gathered from major media outlets like CNBC, The Guardian, Reuters, and CNN, as well as extra datasets from Kaggle. These sources include many headlines and full articles, giving useful information on different topics. Social media data was collected from platforms like Reddit and Twitter, focusing on related posts and titles to show public opinions and trends.

   The data we collected covers the period from 2018 to 2024, which was chosen to include different economic cycles. This period includes important global events and market ups and downs, making the data very useful for analysis. The dataset has about 60,000 entries, including both headlines and full articles. We think this large amount of data provides a strong and varied base for analysis, with the chance to find patterns and insights in different areas.
2) **Stock Price Trends:** We got the actual stock price data from Yahoo, focusing on stock price changes from 2018 to 2024. To predict the trend of the U.S. stock market, we decided to use the S&P 500 index as our main reference. This index tracks the performance of about 80% of the total market value of all publicly traded companies in the U.S., so it is widely recognized as a reliable standard. By using the S&P 500, we can have a good overview of the market's overall performance and make more accurate predictions based on its trends.

B. Preprocessing

The process of text cleaning involves removing unnecessary spaces, symbols, trademarks, and other unwanted elements from the text. This ensures the data is clean and consistent, making it easier to analyze. For date conversion, we will standardize the date column in all datasets to the datetime format. This ensures uniformity and proper organization across the data. Any invalid or incorrect dates will be removed to prevent errors during analysis. Once the data is cleaned and organized, we will merge the news/social media data with the S&P 500 stock data by aligning them based on the date. This step ensures that each piece of news or social media post is matched with the corresponding stock data from the same time period, allowing for accurate comparison and analysis.

C. Selection Feature

To enhance the model's performance, we have decided to incorporate several additional features that can provide valuable insights :

We add some key technical indicators such as the Moving Average (MA), which helps smooth out price data to identify trends, the Relative Strength Index (RSI), which measures the speed and change of price movements to determine overbought or oversold conditions, and the Moving Average Convergence Divergence (MACD), which is used to identify changes in the strength, direction, momentum, and duration of a trend in stock prices.

In addition to these technical indicators, we are including lagging features and volatility features derived from news sentiment. These features are designed to analyze how sentiment in the news might affect the market with a delay. We can analyze how it influences market after a certain lag period.

Finally, we will include lagged price changes and rolling average price changes of the S&P 500 index. These features will serve as additional signals to help predict broader stock market trends. The lagged price changes help capture past movements in the index, while the rolling average smooths out short-term fluctuations, providing a clearer picture of the market' s direction.

D. Evaluation Method

For assessing the performance of our news sentiment analysis to stock price, we employ the following evaluation metrics:

- **Accuracy:** Accuracy is a fundamental metric that measures the overall correctness of the model's predictions. It is the ratio of correctly predicted instances to the total instances.
- **Macro Average:** The Macro Average is the average of the F1 scores across all classes, treating each class equally regardless of its size. It provides a balanced measure of the model's performance by calculating the F1 score for each class individually and then averaging them. This metric is particularly useful when you want to evaluate how well the model performs across different categories without being biased by the size of any specific class.
- **Recall:** Recall, also known as sensitivity or true positive rate, quantifies the model's ability to correctly identify instances of a particular class among all instances of that class.

Together, these metrics give a comprehensive picture of how effectively our model analyzes the relationship between news sentiment and stock price changes, allowing us to make informed adjustments and improvements to its performance.

E. Execution Details (Validation)

In the model validation phase, we use TimeSeriesSplit to partition the training and validation data in a way that respects the temporal sequence of events. This method ensures that the training set always precedes the validation set, which is essential for maintaining the causal relationship inherent in time series data. By using TimeSeriesSplit, we simulate real-world scenarios where past data is used to predict future outcomes.

We apply several machine learning models, including Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), CatBoost, Multi-Layer Perceptron (MLP), and Stacking. Each model undergoes 5-fold cross-validation, ensuring that every instance in the dataset is used for both training and validation. After performing the 5-fold validation, we calculate the average accuracy score for each model. The model that achieves the highest average accuracy is selected as the best-performing model and is then used for subsequent testing.

This process helps ensure that the model selection is based on reliable and consistent performance across multiple folds, allowing us to confidently choose the model that will generalize best to unseen data. Furthermore, by comparing various models, we can identify the most effective algorithm for predicting stock price movements based on news sentiment.

F. Execution Details (Prediction)

In this stage, we select the best-performing model from the previous validation phase, which in this case is XGBoost, to perform the final training and prediction. The data is split into training and testing sets, where the first 80% of the data, covering the years 2018 to 2023, is used to train the model. The remaining 20% of the data, corresponding to the entire year of 2024, is used for testing and making predictions.

Once the model is trained on the 2018-2023 data, it is applied to the 2024 data to predict stock price movements based on news sentiment. To evaluate the model's performance, we use key metrics such as Accuracy and F1 Score. Accuracy helps measure the overall correctness of the predictions, while the F1 Score provides a balanced measure that considers both precision and recall, helping us understand how well the model performs in different situations, especially in cases where the classes are imbalanced.

By analyzing these evaluation metrics, we can assess the effectiveness of the model in predicting stock price trends and its ability to generalize to new data. This final step gives us a clear picture of how well the model captures the relationship between news sentiment and stock price movements, allowing us to make informed decisions and improvements for future predictions.

G. Results

The result of the performance is down below :

The overall accuracy is 53%, slightly higher than random guessing (50%), indicating that the model has some predictive value but still falls short of ideal performance.

Class	Precision	Recall	F1-Score
0 (Decrease)	0.46	0.52	0.49
1 (Increase)	0.60	0.54	0.57
Macro Avg	0.53	0.53	0.53
Weighted Avg	0.54	0.53	0.54
Accuracy	x	x	**0.53**

Performance Metrics

The F1-Score of 0.5725 shows that the model has a slight advantage in balancing precision and recall.

The results suggest that the model performs relatively better in predicting "Uptrend (1)," achieving an F1-Score of 0.57, while its ability to predict "Downtrend (0)" is weaker, with an F1-Score of only 0.49. This indicates that the model is more inclined to capture upward trends but struggles with accuracy and stability when predicting downward trends.

Additionally, the Macro Average scores reflect that the model's performance across different classes is relatively balanced. However, the Weighted Average, which accounts for class imbalances, suggests that the model performs better on classes with a larger proportion of data.

Overall, while the model demonstrates a certain degree of trend prediction capability, it requires further optimization to enhance its ability to capture downward trends and improve its overall accuracy and stability in practical applications.

V. Analysis

Our model's prediction results, as shown in Figure 1, indicate a certain level of correlation between sentiment scores and S&P 500 scores.

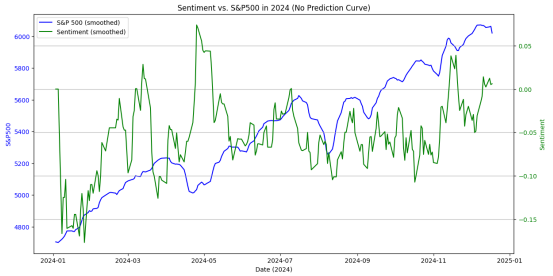

Fig. 1: S&P 500 and sentiment

Summarizing the results of all our models, we observed that the overall accuracy was only 0.5342, while the F1-score was 0.5725. Taking a closer look at the precision for predicting upward and downward trends, the precision for upward trends reached 0.60, whereas the precision for downward trends was 0.46. This indicates that our model performs relatively better in predicting upward trends. We attribute the model's performance to the following reasons:

1. Insufficient Data Volume: During the process of collecting news data, we encountered several challenges and limitations, resulting in a relatively small dataset. To compensate for the lack of data, we incorporated additional data obtained from Kaggle. However, this may have caused an imbalance in the dataset. The limited data volume hindered the model's ability to learn sufficient features, leading to reduced prediction accuracy.

2. Imbalanced Sample Distribution: The dataset may have an imbalanced number of samples for predicting stock price increases or decreases. This imbalance could bias the model towards predicting the majority class. For instance, if the dataset contains more articles with high sentiment scores, possibly due to the authors' preference for writing such articles, this bias might affect the model's performance.

3. Random Walk Theory: The random walk theory posits that asset price changes are entirely random, and future price movements cannot be accurately predicted based on past price changes. According to the Efficient Market Hypothesis, market prices fully reflect all available information, and any new information is immediately reflected in the prices. Since new information is often random and unpredictable, it becomes challenging to achieve returns by analyzing such data. This inherent randomness makes it difficult for our model to analyze the relationship between sentiment scores and S&P 500 values effectively.

To improve our method, we propose the following approaches:

- Expand the sources of data, for example, by gathering information from a wider variety of social media platforms to avoid potential echo chambers on specific platforms.
- For rare classes, apply the SMOTE technique to generate more samples and address the class imbalance issue.
- Incorporate long-term market trends, such as eco-

nomic indicators and inflation rates, as features to mitigate the impact of randomness.

VI. Conclusion

To summarize our study, we utilized various methods, including RF, XGB, LGBM, CatBoost, MLP, and Stacking, to analyze the relationship between the S&P 500 index and sentiment scores. Additionally, we incorporated numerous extra features such as Moving Averages (MA), Relative Strength Index (RSI), and Moving Average Convergence Divergence (MACD), which significantly contributed to our research.

After performing 5-fold validation for each model, the best-performing model was XGBoost, achieving an overall accuracy of 0.5342. The overall F1-score for all models was 0.5725.

Despite our extensive efforts, the models did not perform as well as expected due to three main factors. First, the dataset lacked sufficient samples, preventing the models from learning enough features. Second, the collected data was imbalanced, leading to a bias in the trained models toward predicting the majority classes, thereby reducing accuracy. Third, the limitations inherent in stock price prediction, such as those imposed by the random walk theory, made it challenging to analyze the relationship between the data and stock prices. As a result, the overall performance of the models was suboptimal.

To improve model performance, future work could focus on expanding the dataset, addressing the random walk theory through advanced feature engineering, and exploring more complex model architectures. Continuous optimization in these areas may enhance accuracy and lead to the development of a more reliable stock price prediction system.

Despite facing certain challenges and limitations, this research lays the foundation for further exploration and improvement in *Market Sentiment Analysis in Financial News and Stock Price Prediction*, contributing to the ongoing efforts to enhance the accuracy and robustness of such models.

References

[1] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, Newark, CA, USA, 2019, pp. 205–208, doi: 10.1109/BigDataService.2019.00035.

[2] G. Ranibaran, M.-S. Moin, S. H. Alizadeh, and A. Koochari, "Analyzing Effect of News Polarity on Stock Market Prediction: A Machine Learning Approach," *2021 12th International Conference on Information and Knowledge Technology (IKT)*, Babol, Iran, 2021, pp. 102–106, doi: 10.1109/IKT54664.2021.9685403.

[3] S. K. Bharti, P. Tratiya, and R. K. Gupta, "Stock Market Price Prediction through News Sentiment Analysis & Ensemble Learning," *2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, Gunupur, Odisha, India, 2022, pp. 1–5, doi: 10.1109/iSSSC56467.2022.10051623.