Technology Sector Stock Price Movement Prediction based on Financial News Headlines Sentiment using Dictionary Method and Transformers

Xu Ding
Columbia University
New York, NY
xd2249@columbia.edu

Jordan Israel Ordonez Chaguay
Columbia University
New York, NY
jio2108@columbia.edu

Abstract

Sentiment analysis of financial news headlines has been found to be efficient and helpful in stock price prediction models, which is an important component of a trader's decision-making framework. This paper concentrates on news headlines related to technology sector companies and builds prediction models based on their sentiment scores. Firstly, a list of Tech companies is created based on the Yahoo Finance Screener with filters: in the United States, Mid, Large, and Mega Cap, and positive stock price. Then, average daily sentiment scores of processed financial news headlines dataset are calculated utilizing traditional dictionary methods and transformer-based models including FinBert, DistilBert, Financial-RoBERTa, and SEC-BERT models. Mode numbers of all models' outputs are final sentiment scores and company-level stock price predictions are based on daily average sentiment scores. Finally, the model performance is evaluated among Ridge classifier, ensemble support vector machine, and multi-layer perceptron based on accuracy and F1 score. Since price trends are relatively balanced, Ridge classifier is concluded to be the optimal prediction model based on accuracy metric.

1 Introduction

Economists believed that the market is efficient and stock prices should react rapidly to new public information (e.g., see summary in Fama, 1970). However, the Efficient Market hypothesis assumes the adjustment time of stock price is close to zero and implies that investors cannot gain excess profit from stock market predictions, which contradicts various empirical studies. For example, Lim (2009) concluded that information dissemination barriers and inattention of investors contribute to the delay of stock price adjustment. Therefore, it is possible for investors to gain profits if they make predictions before the market adjusts. During the era of big data, information dissemination barriers are reduced and more investors pay attention to updated information to exploit profits from the stock market. The adjustment time of stock prices is much shorter than decades ago and investors require more technical and systematic approaches to predict market information efficiently.

Before the development of natural language processing, researchers could not directly deal with textual data. For example, Jennings and Starks (1985) analyzed quarterly earnings announcements from companies . Data points were limited and the information content was purely determined by statistics of analysts' forecasts.

The traditional dictionary method was introduced to the finance field and computer scientists collaborated with linguistic experts to develop dictionaries for computers to understand the sentiment of words. Linguistic experts decompose each word into several clusters of emotions and every word could be classified as positive, negative, or neutral. This method can count the number of positive and negative words for each sentence. In the finance domain, if a financial news headline related to company A contains more negative words than positive words, this headline might be a bearish signal of company A. The dictionary method highly depends on the quality of the dictionary and some professional dictionaries are preferable when working on specific tasks. Additionally, the input data must be preprocessed carefully to be compatible with the dictionary.

Another popular approach for sentiment analysis is the transformer-based model. The structure of a transformer can be decomposed into three components: word embedding, encoder and decoder. Word embedding allows the model to capture some contexts within each sentence, which is the largest advantage of transformer compared to the dictionary method.

Word embedding Mikolov et al. (2013) is an unsupervised neural network and can convert each word into a vector representation based on the dataset provided. An encoder can capture both linear and non-linear relationships to appropriately recognize the whole context of a sentence. Using word embedding vectors as input, an encoder is good at text classification assignments. Our model does not include the decoder in the architecture.

After fine-tuning a transformer, this model can label each financial news headline as 0 (negative), 1 (neutral), or 2 (positive). A headline with a higher daily average sentiment score implies that the related company has more positive news today. And daily average sentiment scores can be the independent variable to predict the trend of stock price movement: either up or down. Instead of knowing precise growth rates, it is sufficient for traders to make decisions when they can predict the stock movement direction. The model performance is evaluated among Ridge classifier, support vector machine, and multi-layer perceptron classifier based on accuracy and F1 score.

The structure of this paper is as follows: The next section includes literature reviews of previous work on stock price prediction based on financial news sentiment analysis and recent studies related to transformer applications in the finance domain. The "Data Collection" section explains details of our financial news headline dataset, filters to create our target company list, and historical stock price dataset. The "Sentiment Analysis Methodology" section shows every step of our research project including data pre-processing, transformer model selection, prediction models based on sentiment scores, and model evaluation.

2 Literature Review

Researchers initiated to incorporate sentiment analysis into stock price prediction models after the development of natural language processing techniques. We briefly review recent studies related to financial news sentiment analysis and selection of stock market prediction models in this section.

2.1 Dictionary Method

The first sentiment analysis technique introduced in the finance domain is the dictionary method. Li et al. (2014) utilized sentiment dictionaries manually to analyze financial news articles related to stock market in Hong Kong. Moreover, Xiao and Ihnaini utilized the dictionary method to analyze the sentiment of financial texts in a recent study. (Xiao and Ihnaini, 2023) They used VADER library and sentiment dictionaries to do sentiment analysis. The collection of financial news articles was from Financial PhraseBank and Tweets with financial hashtags.

Both experiment result from Li's and Xiao's papers indicated that the dictionary method improved the predictive accuracy and it would be better to transform polarity into a categorical variable before adding to the model.

2.2 Transformers-based Models

A research team from Google developed an advanced transformer model called Bidirectional Encoder Representations from Transformers (BERT) designed for text classification tasks published in Devlin et al. (2018), which can capture contextual information from both directions and understand the meaning of words better.

The performance of the BERT model depends on the quality of training datasets. Scientists can use datasets from different knowledge domains to develop various BERT-based models.

Due to the extensive training time, researchers consider using fine-tuned BERT models in the project. The most popular finance BERT-based model called FinBERT was published in Araci (2019) and computer scientists continued to tune the model parameters in the following years.

Liu and other researchers did FinBERT sentiment analysis on Tweets posted by users on Stocktwits to forecast the stock price movement: either up or down (Liu et al., 2023). The prediction model achieved the highest accuracy score when using outputs from FinBERT compared with using two-class classification provided by Stocktwits or using outputs from other general sentiment analysis tools.

Researchers also considered other BERT-based models to do sentiment analysis. On one hand, Glodd and Hristova (2023) predicted stock price growth based on sentiment analysis from DistilBERT and concluded that DistilBERT might outperform FinBERT when dealing with company annual financial reports. On the other hand, Gupta and Tayal (2023) implemented RoBERTa to analyze the sentiment of Tweets related to financial keywords and RoBERTa achieved high performance.

Our project aims at choosing the mode number of outputs from the dictionary method and popular BERT-based models to conduct more robust sentiment analysis.

2.3 Stock Market Prediction

After completing the sentiment analysis on financial news headlines, the stock price movement prediction is the final objective. Most recent studies indicate that it is more appropriate to predict the movement direction of stock price instead of the actual stock price growth rate. Decision-making for trading strategies only requires traders to understand the future trend of stock prices. And there are several candidates for the prediction model shown in previous work.

In a recent study, Yenkikar and Babu (2023) compared the performance of several prediction models using sentiment analysis on financial news. They concluded that Ridge classifier and neural network models outperform other popular algorithms based on accuracy and F1 score metrics. Moreover, neural networks were applicable for stock price prediction shown in the paper by BL and BR (2023). Shilpa built the stock price prediction model based on classification neural networks, which can capture non-linear relationships and have good prediction performance for time series data.

Additionally, support vector machine (SVM) is a widely used approach for classification assignments. In Li's paper, the stock price prediction model was the Radial Basis Function (RBF) kernel SVM, which could deal with non-linearly separable data points (Li et al., 2014). Liu utilized another technique called ensemble SVM, which combines the outputs from several SVMs to prevent overfitting problems (Liu et al., 2023). Ensemble SVM trained on several subsets of training data and these subsets were created by sampling with replacement.

Our project concentrates on the company-level prediction of U.S. Technology sector company stocks, which was seldom covered by previous works. And we selected three predictive models recommended by most stock price prediction papers: Ridge classifer, SVM, and multi-layer perceptron neural network classifier.

3 Data Collection

3.1 Financial News Headline Dataset

The dataset is from Kaggle and the author used web-scraping to retrieve financial news headlines from Benzinga.com, a financial news company located in the United States (Aenlle, 2022). The dataset ranging from 2009 to 2020 includes financial news headlines related to 6193 different stocks. Each row of this dataset contains an article headline, a release timestamp of published date, and a stock ticker. Financial analysts from Benzinga provide the stock ticker explaining which stock this news headline relates to.

3.2 Technology Company Selection

Our paper focuses on companies from the technology sector since previous work usually investigates the overall market or companies from other sectors. Yahoo Finance provides a function called Stock Screener, which can filter companies based on specific financial criteria. Stock Screener filters include: Technology sector, in the United States, Mid, Large, and Mega Cap, and positive stock price. Additionally, all the companies must have strong reputations in the industry and at least 200 news headlines in the dataset. Definitions related to sector, market cap, and reputation are available in Yahoo Finance Stock Screener. A list of technology company selections are available in Appendix A.

There are several reasons why these filters have been selected. Benzinga is a company located in the United States and its financial news related to American companies should be more reliable. Additionally, the availability of Yahoo Finance stock price dataset is larger for American companies. Finally, companies with higher market capitalization tend to have more news announcements based on empirical studies. Zhang and Skiena (2010) indicated that there is a strong correlation between market capitalization and total volumes of news each day. Larger firms tend to have more data points available.

After adding the filter of the company list, there are 77922 financial news headlines remaining in our dataset ranging from July 27, 2009 to June 11, 2020, related to 107 technology companies. The list of all companies is available in Appendix B.

3.3 Yahoo Finance Stock Price

Yahoo Finance API called yfinance provides stock prices of all companies in the technology sector and we retrieve the daily Low prices, High prices, Close prices, and adjusted Close prices in our dataset. The reason to exclude Open prices is that most news articles are published after the market opens and when investors observe the headline sentiment score, daily Open prices have been fixed.

Low prices and High prices are important for momentum trading, which means investors buy and sell stocks following the market trend. Momentum traders tend to buy more when stock price increases and sell more when stock price decreases. Their profits mainly depend on values of Low prices and High prices. Portfolio managers pay much attention to Close prices and adjusted Close prices since most stock indices are calculated using Close prices or adjusted Close prices. Quantitative trading strategies must depend on accurate prediction of Close price movement. Therefore, our paper decides to research these stock price movements.

4 Sentiment Analysis Methodology

4.1 News Headlines Data Preprocessing

To decrease the risk of overfitting problems in the transformer models, the news headlines were preprocessed to remove non-English characters, redundant words and meaningless words.

To check for non-English characters, all the uppercase characters transformed into lowercase characters to increase the efficiency of programs. This step is necessary because all sentiment dictionaries and models are designed based on English words.

To determine redundant words, we used word stemming through NLTK. To remove all meaningless words, we used word lists posted by the University of of Notre Dame as these lists cover the common stop words, geographic information, dates, numbers, and human names.

4.2 Dictionary Method Setup

We selected two dictionaries for our research project: Harvard IV-4 Dictionary and Loughran-McDonald Master Dictionary. Both dictionaries label every word positive, neutral, or negative, and provide the ability to calculate the polarity of one financial news headline by counting the number of positive and negative words. For this study, we based polarity on prior work Li et al. (2014) and Glodd and Hristova (2023) where polarity is defined by:

$$\text{polarity} = \frac{\text{positive} - \text{negative}}{\text{positive} + \text{negative}}$$

Loughran-McDonald Master (LM) Dictionary was updated by University of of Notre Dame (2022) and contained word frequency in all 10-K type filings from EDGAR 10-K archive and earnings calls from CapIQ ranging from 1993 to 2021. Since LM Dictionary was created purely based on financial documents, its polarity outputs should be more reliable.

Harvard IV-4 (HIV4) Dictionary was created by Harvard University (2000) and included 11788 common words from different knowledge domains. Some financial news headlines contained some non-professional but emotional words and LM Dictionary automatically treated these words as neutral words. Polarity outputs from HIV4 Dictionary could be a supplement to results from LM Dictionary.

Both dictionaries provided two polarity values for each financial news headline. A hierarchical framework is built to evaluate the overall sentiment score and ran the check condition in the following order:

If both polarity values = 0, this sentence is purely neutral and the overall sentiment score = 1 (neutral).

If LM polarity > 0 and HIV4 polarity > 0, this sentence is purely positive and the overall sentiment score = 2 (positive).

If LM polarity < 0 and HIV4 polarity < 0, this sentence is purely negative and the overall sentiment score = 0 (negative).

If LM polarity > 0 and HIV4 polarity = 0, this sentence is more likely to be positive and the overall sentiment score = 2 (positive).

If LM polarity < 0 and HIV4 polarity = 0, this sentence is more likely to be negative and the overall sentiment score = 0 (negative).

All other sentences are undecided and have no sentiment score.

Based on the above check conditions, the dictionary method successfully labeled 45362 financial news headlines: 8123 negative headlines, 29822 neutral headlines, and 7417 positive headlines as Figure 1 illustrated.

There are more neutral news headlines recognized by the model, which is reasonable for journalism. Outputs from dictionary method can be a benchmark for more advanced sentiment analysis tools.

4.3 BERT-based Models Setup

For this paper, another four BERT-based models are included for sentiment analysis. Dictionary method cannot understand the context of a news headline and thus polarity may not be an accurate metric to evaluate the sentiment score because some words

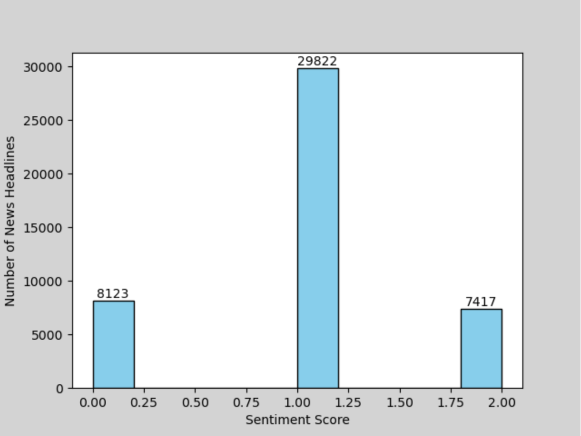


Figure 1: Dictionary Method Sentiment Score

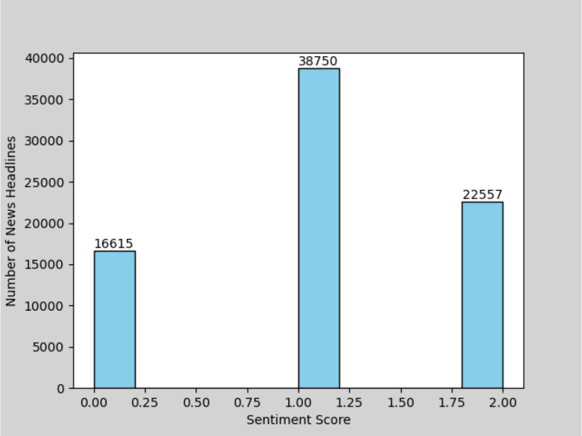


Figure 2: FinBERT Sentiment Score

might contain more emotions than others. For instance, people usually consider the word "blissful" expressing more emotion than the word "satisfied". Therefore, we continued to work on BERT-based models to overcome these limitations for sentiment classification tasks.

4.3.1 FinBERT

BERT only provides a model architecture and training dataset selection is very subjective and flexible. Smart training dataset selection can improve the model performance much. The paper by Araci (2019) first introduced a fine-tuned BERT-based model for the financial domain called FinBERT. Araci suggested using TRC2-financial, Financial PhraseBank, and FiQA Sentiment as training datasets to increase financial text classification accuracy. Since all the training sets are financial documents, FinBERT can perform well in financial sentiment analysis.

Due to the final softmax layer of FinBERT, the outputs were automatically transformed into sentiment scores. FinBERT successfully labeled our 77922 financial news headlines as 0 (negative), 1 (neutral), and 2 (positive) including 16614 negative headlines, 38750 neutral headlines, and 22557 positive headlines as Figure 2 illustrated.

The output pattern is similar to pattern from the dictionary method. There are more neutral news headlines in the dataset and FinBERT captures some additional positive headlines due to its ability to understand the context of sentences.

4.3.2 SEC-BERT

The model architecture of SEC-BERT is very similar to FinBERT but SEC-BERT selects the U.S. Securities and Exchange Commission (SEC) 10-K filing database ranging from 1993 to 2019 as the training dataset. Therefore, SEC-BERT might understand U.S. financial news headlines better and the training dataset contains sufficient stock market knowledge.

Same as FinBERT, SEC-BERT provided 3 different classification outputs and generated sentiment scores: 0 (bearish), 1 (neutral), and 2 (bullish). SEC-BERT to labeled 77922 financial news headlines: 10657 negative headlines, 49215 neutral headlines, and 18050 positive headlines as Figure 3 illustrated.

SEC-BERT is conservative in sentiment analysis compared to FinBERT and labeled more headlines as neutral because its training data contains more official accounting documents, which is a new financial perspective to analyze the sentiment.

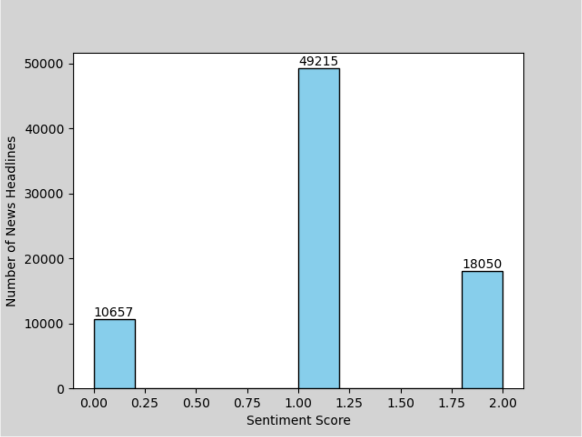


Figure 3: SEC-BERT Sentiment Score

4.3.3 DistilBERT

Although fine-tuned BERT models save researchers much time to adjust model parameters, BERT-based models take longer than traditional machine learning models. Sanh and his research team invented a new BERT architecture called DistilBERT (Sanh et al., 2019). DistilBERT has better optimization algorithm and can be a supplement to FinBERT because DistilBERT is more robust to outliers.

There was no difference in training dataset and outputs were categorized in the final layer. DistilBERT successfully labeled our 77922 financial news headlines as 0 (negative), 1 (neutral), and 2 (positive) including 15642 negative headlines, 37940 neutral headlines, and 24340 positive headlines. Outputs from DistilBERT are similar to results from FinBERT as Figure 4 illustrated.

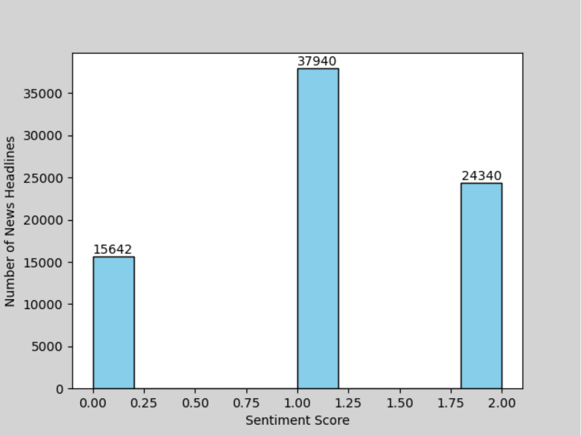


Figure 4: DistilBERT Sentiment Score

4.3.4 RoBERTa

Soleimanian investigated a more efficient version of BERT and invented DistilBERT. Liu hoped to discover a more robust and accurate version of BERT and he created an updated version model called RoBERTa. He adjusted parameters of the original BERT model and allowed the model to study the training dataset more carefully. Training dataset of financial RoBERTa is much larger than FinBERT and includes Corporate Social Responsibility (CSR) reports and Environmental, Social, and Governance (ESG) News in addition to the original dataset. Therefore, adding RoBERTa into the sentiment analysis framework can increase the robustness of results.

RoBERTa successfully labeled our 77922 financial news headlines as 0 (negative), 1 (neutral), and 2 (positive) including 18148 negative headlines, 25897 neutral headlines, and 33877 positive headlines as Figure 5 illustrated.

There were more positive headlines captured by RoBERTa compared to other BERT-based models. RoBERTa might analyze financial news headlines from social responsibility perspectives, and some neutral headlines related to environmental protection are labeled as positive sentiment.

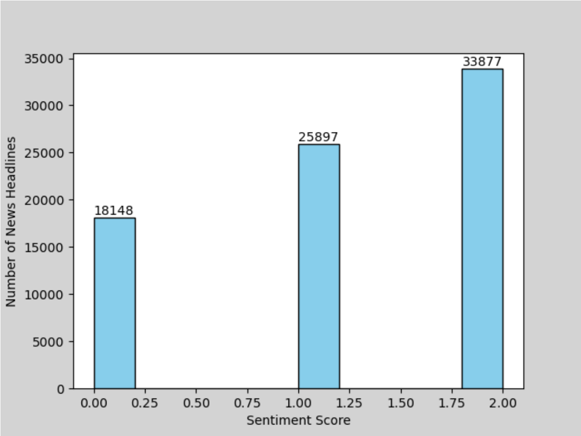


Figure 5: RoBERTa Sentiment Score

4.4 Daily Average Sentiment Score

Five different NLP models were included for sentiment analysis in our research project and each of them provided a sentiment score for every financial news headline. If there were multiple news articles available for one company, we took the daily average sentiment score since the impact from good or bad financial news was not incremental within one day and let all the models vote for the overall sentiment score, which was the mode among five model outputs.

Using the mode number can generate more robust sentiment scores and minimize influence from outliers. For each company, there is a time series of daily average sentiment score and the collection of all companies' data is shown in Figure 6.

5 Prediction Methodology

All the daily Low prices, High prices, Close prices, and adjusted Close prices are numerical variables and the price trends are stock price daily growth rates. As mentioned in "Introduction" section, since traders do not need a specific growth rate for the decision-making procedure, daily growth rates are transformed into categorical versions. If growth rate > 0, the price trend is labeled as 1 (up). If growth rate < 0, the price trend is labeled as -1 (down).

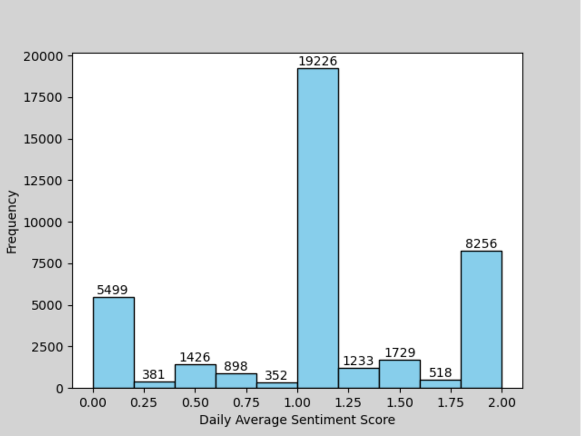
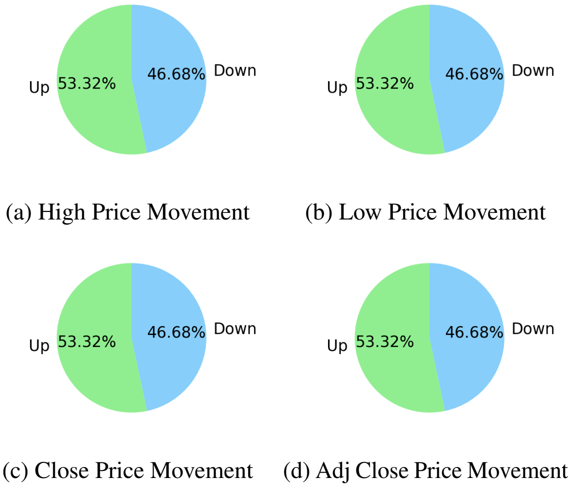


Figure 6: Daily Average Sentiment Score of All Companies

Different traders prefer predictions of different stock prices due to the logic of their trading strategies. Momentum traders focus more on Low and High price trend prediction because they can gain profits if the difference between these two prices is large. Quantitative traders hope to minimize trading risk and they should track Close and adjusted Close price trends carefully. Therefore, price movement predictions for four different prices are all important for financial industry. The prediction for each company is independent based on previous work (Liu et al., 2023). Empirical studies show that stock returns can also be influenced by personal preference for individual stock and unexpected company-specific events. Therefore, company-level prediction performs better than sector-level prediction.



(a) High Price Movement

(b) Low Price Movement

(c) Close Price Movement

(d) Adj Close Price Movement

Figure 7: Stock Price Movement

Figure 7 indicates that up and down trends are relatively balanced for High, Low, Close, and Adjusted Close prices, which means that up and down trends are equally represented by the dataset. Therefore, there is no unbalanced data problem for the prediction model.

5.1 Variable Description

There are two predictor variables: daily average sentiment score and previous daily average sentiment score. Both variables are continuous and ranging from 0 to 2. According to empirical studies Lim (2009), information from previous period might influence the stock price movement during current period. Including previous daily average sentiment score can decrease the risk of omitted-variable bias in prediction.

Four predicted variables are categorical High, Low, Close, adjusted Close price daily growth rate. All variables are binary and can be either 1 (up trend) or -1 (down trend).

5.2 Cross Validation

Cross validation was applied for the prediction model training to increase the robustness of results. Time series data requires different steps for cross validation because random shuffle is not allowed. We utilized a rolling window approach mentioned in previous work (Liu et al., 2023).

5.3 Prediction Models

There are three different prediction models included in the research project and they are all classification models because predictor variables are binary. The model selection depends on prior work by BL and BR (2023) and Liu et al. (2023).

Ridge classifier is an efficient regularized linear model, which is robust to outliers.

Ensemble support vector machine with bagging can find the optimal hyperplane for classification and perform well with a relatively small dataset. And this model is less likely to have overfitting problems.

Multi-layer perceptron classifier can capture non-linear relationships and it is a shallow neural network, which does not require a large training dataset.

6 Model Evaluation

The model evaluation follows the framework in Li et al. (2014). Traders often work on stock portfolios including various stocks and their trading profits are less influenced by an individual stock price prediction. This evaluation method can find which model has better prediction performance on more stock prices among all 107 technology sector companies.

The model evaluation is available for High price prediction in Table 1, for Low price prediction in Table 2, for Close price prediction in Table 3, and for Adjusted Close price prediction in Table 4.

Based on the accuracy metric, Ridge classifier outperforms two other models in all four types of price prediction. But based on the F1 score metric, multi-layer perceptron classifier has the highest performance.

	Ridge	SVM	Perceptron
Accuracy	48	21	38
F1 Score	36	25	46

Table 1: High Price Movement Prediction

	Ridge	SVM	Perceptron
Accuracy	55	20	32
F1 Score	42	18	47

Table 2: Low Price Movement Prediction

	Ridge	SVM	Perceptron
Accuracy	49	19	39
F1 Score	39	20	48

Table 3: Close Price Movement Prediction

	Ridge	SVM	Perceptron
Accuracy	51	20	36
F1 Score	41	18	48

Table 4: Adjusted Close Price Movement Prediction

7 Conclusions

In this paper, we utilized the dictionary method and BERT-based models to do sentiment analysis on financial news headlines. Most models recognized more neutral headlines in the dataset, which was similar to the pattern of daily average sentiment scores. It is reasonable for financial journalists to publish more neutral news articles. However, since our dataset is downloaded from Kaggle, the data quality and bias might be a concern and some potential solutions are discussed in the Future Works section.

Moreover, our paper explored two distinct metrics to investigate the optimal prediction model: the accuracy and F1 score. Accuracy is a more appropriate metric if the dataset is balanced; otherwise, F1 score is better for an unbalanced dataset. Our dataset is relatively balanced for price trends as mentioned in Prediction Methodology section.

Therefore, the model selection can depend on the accuracy metric and Ridge classifier might be the most optimal model for stock price movement prediction. However, SVM does not perform well even if we add a non-linear kernel function to the model. The reason might be that SVM is less robust to outliers or the data characteristics are not suitable for SVM models.

8 Future Works

First of all, the dataset quality can be improved if more news sources are included. Bloomberg is the most comprehensive news vendor in the market and it indicates the popularity and relevance of each news article. However, retrieving information from Bloomberg requires more time and research members, which should be a long-term research goal. Additionally, Tweets with financial hashtags might be a potential data source if obtaining sufficient funding for Twitter API as mentioned in Xiao and Ihnaini (2023).

Secondly, this paper only focuses on technology sector companies and there are still several sectors not covered by prior works, such as healthcare and financial sectors. We can continue to work on stock price movement predictions in these sectors to expand the investment universe for traders. If more sectors are included, we can work on sector-level stock price trend predictions.

Thirdly, a multi-layer perceptron classifier might achieve higher score in accuracy metric if implementing sufficient parameter tuning. Some hyperparameters can improve the model performance and our cross-validation step can minimize overfitting problems. Future works can investigate the optimal parameters for multi-layer perceptron classifier following the framework published by Schilling et al. (2015).

Moreover, our research was limited to performing sentiment analysis and stock price movement prediction on companies that were U.S.-based and had financial news headlines published in the English language. Expanding into non-U.S.-based companies and gathering non-English text headlines could potentially aid in building more robust price movement prediction models. Future works could look into multilingual sentiment analysis by utilizing transformer models fine-tuning on non-English text documents.

Finally, causal inference between daily average sentiment scores and stock price movements is an interesting problem, which is still controversial in the industry and academic institutions. We can try instrumental variables or propensity score matching techniques in future works to verify the causal relationship between sentiment scores and stock price trends.

References

Miguel Aenlle. 2022. Daily financial news for 6000+ stocks. Accessed November 17, 2023.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Shilpa BL and Shambhavi BR. 2023. Combined deep learning classifiers for stock market prediction: integrating stock price and news sentiments. *Kybernetes*, 52(3):748–773.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eugene F Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Alexander Glodd and Diana Hristova. 2023. Extraction of forward-looking financial information for stock price prediction from annual reports using nlp techniques.

Aditya Gupta and Vijay Kumar Tayal. 2023. Analysis of twitter sentiment to predict financial trends. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pages 1027–1031. IEEE.

Robert Jennings and Laura Starks. 1985. Information content and the speed of stock price adjustment. *Journal of Accounting Research*, 23(1):336–350.

Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Kian-Ping Lim. 2009. The speed of stock price adjustment to market-wide information. *Available at SSRN 1412231*.

Jin-Xian Liu, Jenq-Shiou Leu, and Stefan Holst. 2023. Stock price movement prediction based on stocktwits investor sentiment using finbert and ensemble svm. *PeerJ Computer Science*, 9:e1403.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

University of Notre Dame. Stopwords lists. Accessed November 17, 2023.

University of Notre Dame. 2022. Loughran-mcdonald master dictionary. Accessed November 17, 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Nicolas Schilling, Martin Wistuba, Lucas Drumond, and Lars Schmidt-Thieme. 2015. Hyperparameter optimization with factorized multilayer perceptrons. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II 15*, pages 87–103. Springer.

Mohammad Soleimanian. Financial-roberta. Accessed November 17, 2023.

Harvard University. 2000. Harvard iv-4 dictionary. Accessed November 17, 2023.

Qianyi Xiao and Baha Ihnaini. 2023. Stock trend prediction using sentiment analysis. *PeerJ Computer Science*, 9:e1293.

Anuradha Yenkikar and C Narendra Babu. 2023. Comparison of machine learning algorithm for stock price prediction using sentiment analysis. In *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–6. IEEE.

Wenbin Zhang and Steven Skiena. 2010. Trading strategies to exploit blog and news sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 375–378.

A Group Contributions

We both worked on the Prediction Methodology and Future Works sections.

Xu focused on financial news headlines preprocessing, sentiment analysis tools, and model evaluation. Xu wrote the Introduction, Literature Review, Sentiment Analysis Methodology, and Model Evaluation sections.

Jordan worked on retrieving stock prices and prediction model. Jordan wrote the Data Collection and Conclusions sections.

B Technology Company List

Stock Ticker	Company Name
AAPL	Apple Inc.
ACIW	ACI Worldwide Inc.
ACN	Accenture plc
ADBE	Adobe Inc.
ADI	Analog Devices Inc.
ADSK	Autodesk Inc.
AMAT	Applied Materials Inc.
AMD	Advanced Micro Devices Inc.
APH	Amphenol Corporation
ARW	Arrow Electronics Inc.
ASML	ASML Holding NV
AVGO	Broadcom Inc.
AVT	Avnet Inc.
AZPN	Aspen Technology Inc.
BDC	Belden Inc.
BMI	Badger Meter Inc.
BR	Broadridge Financial Solutions Inc.
CACI	CACI International Inc.
CALX	Calix Inc.
CAMT	Camtek Ltd.
CDNS	Cadence Design Systems Inc.
CHKP	Check Point Software Technologies Ltd.
CIEN	Ciena Corporation
COHR	Coherent Inc.
CRM	Salesforce.com Inc.
CRUS	Cirrus Logic Inc.
CSCO	Cisco Systems Inc.
CTSH	Cognizant Technology Solutions Corporation
CVLT	Commvault Systems Inc.
DIOD	Diodes Incorporated
DOX	Amdocs Limited
ENTG	Entegris Inc.
ERIC	Telefonaktiebolaget LM Ericsson
EXLS	ExlService Holdings Inc.
EXTR	Extreme Networks Inc.
FICO	Fair Isaac Corporation
FIS	Fidelity National Information Services Inc.
FLEX	Flex Ltd.
FLT	Fleetcor Technologies Inc.
FN	Fabrinet
FORM	FormFactor Inc.
FSLR	First Solar Inc.
FTNT	Fortinet Inc.
G	Genpact Limited
GRMN	Garmin Ltd.
IBM	International Business Machines Corporation
INFY	Infosys Limited
INTU	Intuit Inc.
IPGP	IPG Photonics Corporation
IT	Gartner Inc.
ITRI	Itron Inc.
JBL	Jabil Inc.
JNPR	Juniper Networks Inc.
KEYS	Keysight Technologies Inc.
KLAC	KLA Corporation
LDOS	Leidos Holdings Inc.
LFUS	Littelfuse Inc.
LOGI	Logitech International S.A.
LPL	LG Display Co. Ltd.
LRCX	Lam Research Corporation
LSCC	Lattice Semiconductor Corporation
MANH	Manhattan Associates Inc.
MCHP	Microchip Technology Inc.
MKSI	MKS Instruments Inc.
MPWR	Monolithic Power Systems Inc.
MRVL	Marvell Technology Group Ltd.
MSI	Motorola Solutions Inc.
MU	Micron Technology Inc.
NICE	NICE Ltd.
NOK	Nokia Corporation
NTAP	NetApp Inc.
NVDA	NVIDIA Corporation
NXPI	NXP Semiconductors N.V.
OLED	Universal Display Corporation
ORCL	Oracle Corporation
OTEX	Open Text Corporation
PLXS	Plexus Corp.
POWI	Power Integrations Inc.
PRGS	Progress Software Corporation
QCOM	Qualcomm Incorporated
QRVO	Qorvo Inc.
RMBS	Rambus Inc.
ROG	Rogers Corporation
SANM	Sanmina Corporation
SNPS	Synopsys Inc.
SNX	Synnex Corporation
ST	Sensata Technologies Holding plc
STM	STMicroelectronics N.V.

Stock Ticker	**Company Name**
STX	Seagate Technology Holdings plc
SYNA	Synaptics Incorporated
TDC	Teradata Corporation
TEL	TE Connectivity Ltd.
TRMB	Trimble Inc.
TSEM	Tower Semiconductor Ltd.
TSM	Taiwan Semiconductor Manufacturing Company Limited
TXN	Texas Instruments Incorporated
TYL	Tyler Technologies Inc.
UMC	United Microelectronics Corporation
VMW	VMware Inc.
VSAT	ViaSat Inc.
VSH	Vishay Intertechnology Inc.
WDC	Western Digital Corporation
WEX	WEX Inc.
WIT	Wipro Limited
WNS	WNS (Holdings) Limited
XRX	Xerox Holdings Corporation
ZBRA	Zebra Technologies Corporation