Dean&Francis
ISSN 2959-6157

Stock Price Trend Prediction Based on News Headlines

Linxu Dai[1, *]

[1]School of Electronic and Information Engineering, Tongji University, Shanghai, China
*Corresponding author: dailinxu@tongji.edu.cn

Abstract:

Behavioral finance shows market sentiment affects market trends. And news headlines, as information carriers with emotional elements. Analyzing the day's trending news headlines enables the prediction of the day's market trends. This study focuses on using news headlines to forecast stock price trends, aiming to capture sentiment's impact on overall stock price changes macroscopically. It uses 25 most viewed Yahoo Finance news headlines (2000–2016) as text data, combined with corresponding Dow Jones Industrial Average (DJIA) fluctuations. Algorithms like Random Forest, XGBoost, Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) were used to learn the mapping between headline features and stock price movements for prediction. Among these models, Logistic Regression performed best: highest accuracy (85.71%), most balanced predictive ability for rises/falls (0.01 F1-score difference), and fast training speed (1.26 seconds). And all models achieved over 83% accuracy, verifying news headlines' sentiment value for stock price prediction and providing a reference for further research.

Keywords: News Headlines; Stock Price Prediction; Machine Learning; Natural Language Processing.

1. Introduction

In today's era of advanced informatization and digitalization, the complexity of financial markets is growing steadily. As one of the core research topics in the financial field, stock price trend prediction has always attracted widespread attention from researchers.

Traditional stock price prediction methods often conduct analysis based on historical price data and basic information such as corporate financial statements. Many studies have improved the structure of Recurrent Neural Networks (RNNs) to enhance prediction accuracy. For instance, a hybrid model combining Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) is used to predict the historical time-series data of stocks [1]; alternatively, an LSTM model improved with an attention mechanism is applied to forecast stock price trends and formulate trading strategies [2]. Additionally, Stacked GRU demonstrates outstanding performance in capturing long-term market trends through multi-layer feature extraction, and is particularly suitable for addressing the non-stationary characteristics of financial data [3]. Some other studies choose to improve the Trans-

former architecture. For example, the MASTER model (Market-Aware Stock Transformer) can effectively handle the correlation between stocks by alternately aggregating information within and between stocks, enabling more accurate predictions [4]. Another study proposes a Kernel-based Hybrid Interpretable Transformer (KHIT) model, which incorporates a new loss function and converts high-frequency stock price prediction from a classification task to a time-series regression task to address the prediction challenge in non-stationary stock markets [5].

While these aforementioned methods have accumulated rich experience through long-term practice, their limitations have become increasingly apparent when facing an ever more complex and volatile market environment—especially since these approaches rely solely on historical price data and corporate financial statements. In fact, stock prices are influenced by multiple factors such as the economic landscape, political developments, market sentiment, and corporate operating conditions; thus, it is challenging to make predictions based solely on historical stock prices.

Stock price prediction can also be achieved through the direction of sentiment analysis: by collecting content such as news data from the internet, and then using methods like machine learning to analyze the emotional tendency of this data, stock prices can be predicted. For example, combining news sentiment analysis with the DBSCAN clustering algorithm to construct a sentiment index based on text polarity has improved prediction accuracy [6].

This paper presents a stock price trend prediction method based on news data, which aims to predict the overall stock price trends of the entire market from a more macro perspective. News typically reflects the general views of most people on the current market situation and also exerts a significant impact on market performance. News headlines, in turn, can basically summarize the main content of articles or reflect the overall emotional tendency of the articles. Therefore, in this experiment, natural language processing technology is used to conduct sentiment analysis on news headlines. Combined with the day's stock price trends, the Random Forest algorithm is applied to derive the mapping relationship between headline content and stock price trends, thereby enabling prediction.

2. Dataset

2.1 Data Sources

The dataset used in this experiment is a combined set of news headlines and stock price movement trends. Data collection spanned from 2000 to 2016. The data frame includes 25 columns, and each corresponding to one of the 25 most viewed news headlines of a given day, a Date column, and a Label column—where a label of 1 indicates a stock price increase and 0 indicates a decrease. The news data was sourced from Yahoo Finance, while the labels were derived from the upward or downward movements of the U.S. Dow Jones Industrial Average (DJIA), which reflects the overall market's stock price changes. The specific format of the dataset is presented in Table 1, with a subset of the data included for demonstration purposes.

Table 1. Sample of the Dataset

Date	Label	Top1	Top2	Top3	Top4	Top5
2000/1/3	0	A 'hindrance to operations': extracts from the leaked reports	Scorecard	Hughes' instant hit buoys Blues	Jack gets his skates on at ice-cold Alex	Chaos as Maracana builds up for United
2000/1/4	0	Scorecard	The best lake scene	Leader: German sleaze inquiry	Cheerio, boyo	The main recommendations
2000/1/5	0	Coventry caught on counter by Flo	United's rivals on the road to Rio	Thatcher issues defence before trial by video	Police help Smith lay down the law at Everton	Tale of Trautmann bears two more retellings
2000/1/6	1	Pilgrim knows how to progress	Thatcher facing ban	McIlroy calls for Irish fighting spirit	Leicester bin stadium blueprint	United braced for Mexican wave
2000/1/7	1	Hitches and Horlocks	Beckham off but United survive	Breast cancer screening	Alan Parker	Guardian readers: are you all whingers?

ISSN 2959-6157

2.2 Data Processing

2.2.1 Dataset Splitting

The dataset was split into a training set and a test set based on the "Date" column in the dataset. Specifically, data from 2000 to the end of 2014 was used as the training set, while data from the start of 2015 to the end of 2016 served as the test set. This splitting strategy facilitates subsequent model training and result evaluation.

2.2.2 Text Preprocessing

First, all non-alphabetic characters (such as numbers, punctuation marks, and special symbols) in the news headlines were removed, leaving only alphabetic characters to avoid interference from irrelevant characters in the analysis. Second, all text was converted to lowercase to eliminate interference caused by case differences (e.g., preventing "Stock" and "stock" from being treated as two distinct words). Finally, the 25 headlines from each day were merged into a single string to increase information volume and more comprehensively reflect the daily news context.

2.2.3 Feature Extraction

The core of converting text features into numerical features is to transform the text collection into a word frequency-based numerical matrix. First, all preprocessed texts are traversed, and a unique index is assigned to each distinct word to generate a vocabulary. Then, each text is converted into a vector based on this vocabulary—where each element in the vector represents the frequency of the corresponding word in that text—ultimately forming a feature matrix. Additionally, to preserve the correlational relationships between words, the experiment adopts bigrams for feature extraction. Instead of using individual words as features, it takes two consecutive words as a single feature, which better aligns with human language patterns.

3. Research Methodology

3.1 Model Training

The models take the "bigram feature matrix" as input and stock price movements (rise/fall) as output. Algorithms including Random Forest, XGBoost, Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) are used to learn the mapping relationship. In this process, the algorithms analyze which bigram features are highly correlated with "price increase" (Label = 1) or "price decrease" (Label = 0). For instance, the models may identify that when terms like "profit growth" appear, stocks are more likely to rise; whereas when terms such as "loss" emerge, stocks are more prone to fall.

3.2 Prediction and Evaluation

The news data from the test set is input into the trained models to predict stock price trends. The predicted trends are then compared with the actual stock price trends in the test set, and the prediction performance of the models is evaluated

3.3 Evaluation Metrics

This experiment evaluates the prediction results from the following aspects:

3.3.1 Accuracy

It refers to the proportion of dates where the model's prediction (of a price fall/rise) matches the actual price movement. Accuracy is suitable for measuring the overall predictive ability of a model, especially in scenarios with balanced data (e.g., the number of positive and negative samples is roughly equal). However, overemphasis on accuracy can render the model completely ineffective in imbalanced data scenarios. For example, if 90% of a stock dataset consists of "price rise samples (positive cases)", the model can achieve 90% accuracy simply by predicting "rise" for all cases—yet its ability to predict "price fall samples (negative cases)" will be zero. In such cases, accuracy fails to reflect the model's true value.

3.3.2 Recall

It represents the proportion of dates with actual price falls/rises that are correctly predicted as such by the model. Recall measures the model's ability to avoid missing positive cases, focusing on whether true positive cases are correctly identified. Nevertheless, overemphasis on recall may lead to an increase in false positives. For instance, to avoid missing any stocks that will rise (improving recall), the model might misclassify more falling stocks as rising, which increases investment risks.

3.3.3 F1-Score

As the harmonic mean of accuracy and recall, it balances the trade-off between the two metrics and provides a comprehensive evaluation of model performance.

In addition, the training time of each model is recorded in this experiment to assess the model's training speed and compare the efficiency of different models.

4. Experimental Results

The comprehensive prediction accuracy, precision, recall, and F1-score for price fall predictions, precision, recall, and F1-score for price rise predictions, as well as the

training time of the five models are presented in Table 2.

Table 2. Performance Comparison of Five Prediction Models

		Prediction of Falling (0)			Prediction of Rising (1)			
Model Name	General Accuracy	Accuracy	Recall	F1-Score	Accuracy	Recall	F1-Score	training time(s)
Random Forest	85.45%	0.95	0.75	0.83	0.80	0.96	0.87	10.9071
Logistic Regression	85.71%	0.85	0.85	0.85	0.86	0.86	0.86	1.2642
Naive Bayes	84.66%	0.93	0.74	0.83	0.79	0.95	0.86	0.0273
Support Vector Machine	84.66%	0.83	0.86	0.85	0.86	0.83	0.85	0.1172
XGBoost	83.86%	0.83	0.84	0.84	0.84	0.84	0.84	13.6215

The following conclusions can be drawn from this experiment:

First, the comprehensive prediction accuracy of all five models exceeds 83%, reaching a generally high level. This indicates that news headlines contain sufficient market sentiment information and can serve as an effective reference for stock trend prediction.

Among the five models, Logistic Regression demonstrates the optimal comprehensive performance. It achieves the highest accuracy (85.71%) and exhibits balanced ability in identifying both price falls (Label=0) and price rises (Label=1) and the difference in its F1-scores for the two categories is merely 0.01, with no obvious bias. Thus, it is well-suited for scenarios requiring balanced judgment of price rises and falls.

Random Forest and Naive Bayes tend to prioritize capturing price rises: both models achieve extremely high recall rates for Label 1 (price rises) at 96% and 95% respectively, meaning they can identify nearly all actual price rise cases. However, their recall rates for Label 0 (price falls) are relatively low at 75% and 74% respectively, indicating that these models are prone to misclassifying price falls as price rises. On the other hand, this characteristic makes them suitable for scenarios where it is better to misclassify non-rising cases as rising than to miss actual rising cases, such as investment decisions involving high risk tolerance.

Support Vector Machine (SVM) is characterized by good balance but slightly lower accuracy: its F1-scores for both Label 0 (price falls) and Label 1 (price rises) are 0.85, showing a balance close to that of Logistic Regression. However, its overall accuracy stands at 84.66%, which is relatively lower.

In contrast, XGBoost performs the worst, with the lowest accuracy (83.86%) and no advantages in identifying either label category. This underperformance is attributed to unoptimized parameters—for instance, parameters such as tree depth and learning rate require further tuning.

In terms of training time, Naive Bayes is the fastest, taking only 0.0273 seconds—this duration is almost negligible. The core reason for its speed lies in its characteristic of not requiring parameter optimization; it only needs to calculate probabilities, thus completely avoiding complex iterative solving processes; Support Vector Machine (SVM) is also fast, with a training time of 0.1172 seconds. Its speed is mainly attributed to the model's avoidance of the high computational cost of non-linear kernels, relying instead on efficient linear solving; Logistic Regression also has a short training time (1.2642 seconds). This is because it has a linear structure, its gradient function converges quickly, and no nested calculations are needed during the iteration process—all of which significantly improve its efficiency; Random Forest takes 10.9071 seconds. The key reason is the large number of trees in the model: each tree needs to traverse possible split points of features. Additionally, the text data in this experiment has a high feature dimension, which further increases the computational load of split judgment. The accumulation of these computations leads to a longer training time. However, Random Forest has the advantage that trees are independent of each other and can be trained in parallel, which reduces the computational time to a certain extent; XGBoost has the longest training time at 13.6215 seconds. This is due to its serial iteration (i.e., each tree depends on the results of the previous tree), as well as its adoption of complex tree structure optimization and intricate splitting strategies—these factors result in high computational costs, making its training speed much slower than linear

ISSN 2959-6157

models.

5. Discussion

This experiment has completed a stock price trend prediction based on news headlines and conducted a comparative analysis of the results from several models. However, there are multiple aspects that can be improved or further explored in future research.

In terms of the dataset, instead of using only news headlines for model training, collecting full news articles would allow for more comprehensive capture of the daily news sentiment by increasing the volume of training data. It can enhance the reliability of the model's predictions.

Exploring text data from other sources is also a viable direction. For example, Twitter user data—including tweets related to globally renowned enterprises—could be adopted to predict stock prices [7].

In the data preprocessing stage, the original program simply merges the 25 headlines. However, the 25 headlines actually differ in importance: the higher a headline ranks, the greater its importance. Therefore, different weights can be added before merging, and headlines with higher view counts can be assigned greater weights.

In the feature extraction stage, the Bag-of-Words model can be replaced with TF-IDF. Term Frequency (TF) refers to the frequency of a word in a single document, reflecting the importance of the word in the current document; Inverse Document Frequency (IDF) measures the commonness of the word in the entire corpus, which can highlight words with higher importance in news headlines [8]. Additionally, a financial domain sentiment lexicon can be introduced to establish correspondences between semantically similar words and the subset of financial sentiment words generated by human experts [9]. This helps capture vocabulary specific to the financial domain and improve prediction accuracy.

In terms of model training, model parameters can also be adjusted to improve accuracy. For example, the number of trees used in the Random Forest model can be increased or decreased, and parameters of XGBoost, such as tree depth and learning rate can be adjusted to achieve the optimal model training effect.

Lastly, the prediction of this experiment only focuses on upward or downward trends. If news data can be combined with the actual stock prices to directly predict stock prices rather than trends, the model will have greater practical significance. For instance, large language models can be used to analyze multimodal data—including news reports, historical stock prices, and a company's financial report data—to predict stock price fluctuations [10].

6. Conclusion

This experiment used machine learning to explore the mapping between news headline features and stock price movements, achieving stock trend prediction and comparing five models' performance.

Results showed all models had a comprehensive accuracy exceeding 83%: Logistic Regression performed best, ranking first with 85.71% accuracy and strong balance (F1-scores were 0.85 and 0.86 respectively); Random Forest and Naive Bayes prioritized capturing rises (rise recall: 96%, 95%) but had low fall recall (75%, 74%), suitable for high-risk-tolerance scenarios; Support Vector Machine (SVM) had similar balance to Logistic Regression (F1: 0.85 for both) but lower accuracy (84.66%); XGBoost performed worst (accuracy: 83.86%), presumably due to unoptimized parameters.

The experiment also verified the reference value of market sentiment in news headlines for stock prediction, supporting future research. Future improvements include: expanding data scope and adding weights; optimizing feature extraction with TF-IDF and financial sentiment lexicons; tuning model parameters; and combining multimodal data to predict stock prices directly via large language models, enhancing practical value.

References

[1] Hossain M A, Karim R, Thulasiram R, Bruce N D B and Wang Y, Hybrid Deep Learning Model for Stock Price Prediction, 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 2018, pp. 1837-1844, doi: 10.1109/SSCI.2018.8628641.

[2] Cheng L C, Huang Y H and Wu M E, Applied attention-based LSTM neural networks in stock prediction, 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4716-4718, doi: 10.1109/BigData.2018.8622541.

[3] Melyani N A, Angraini M and Afdal M, Application of Bidirectional Gated Recurrent Unit and Stacked Gated Recurrent Unit Algorithms for Stock Price Prediction, 2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS), Bandung, Indonesia, 2025, pp. 1-6, doi: 10.1109/ICADEIS65852.2025.10933033.

[4] Li T, Liu Z, Shen Y, Wang X, Chen H, & Huang S. MASTER: Market-Guided Stock Transformer for Stock Price Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 38(1), 162-170. 2024. https://doi.org/10.1609/aaai.v38i1.27767

[5] Lin F, Li P, Lin Y, Chen Z, You H and Feng S, Kernel-based Hybrid Interpretable Transformer for High-frequency Stock Movement Prediction, 2022 IEEE International Conference on

Data Mining (ICDM), Orlando, FL, USA, 2022, pp. 241-250, doi: 10.1109/ICDM54844.2022.00034.
[6] Parvatha L S, Naga Veera Tarun D, Yeswanth M and Kiran J S, Stock Market Prediction Using Sentiment Analysis and Incremental Clustering Approaches, 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023, pp. 888-893, doi: 10.1109/ICACCS57279.2023.10112768.
[7] Harguem S et al., Machine Learning Based Prediction of Stock Exchange on NASDAQ 100: A Twitter Mining Approach, 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2022, pp. 01-10, doi: 10.1109/ICCR56254.2022.9996008.
[8] Martineau J, & Finin T. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. Proceedings of the International AAAI Conference on Web and Social Media, 3(1), 258-261. 2009. https://doi.org/10.1609/icwsm.v3i1.13979
[9] Du Z J, Huang A G, Wermers R, Wu W F, Language and Domain Specificity: A Chinese Financial Sentiment Dictionary, Review of Finance, Volume 26, Issue 3, May 2022, Pages 673–719, https://doi.org/10.1093/rof/rfab036
[10] Elahi A and Taghvaei F, Combining Financial Data and News Articles for Stock Price Movement Prediction Using Large Language Models, 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 4875-4883, doi: 10.1109/BigData62323.2024.10825449.