Enhancing Stock Price Forecasting with Social Media Text: A Comparative Study of Machine Learning Approaches

Bingchang Li[a]
Faculty of Business and Management, Beijing Normal University-Hong Kong
Baptist University United International College, Tangjiawan Town, Zhuhai, China

Keywords: Stock Price Forecasting, Machine Learning, Social Media, Nature Language Processing.

Abstract: With the power to influence investment decisions and provide market stability, stock price forecasting is essential to financial research. The use of machine learning techniques for prediction has gained popularity as technology has progressed. Researchers have proposed incorporating social media textual data to improve prediction accuracy. However, the efficacy of this strategy is still debatable because different kinds of textual information might produce varied results. Certain texts cause predictability to rise dramatically, while others cause it to fall. This research explores the use of machine learning techniques to predict stock prices using text data from social media platforms. Using Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) text representations, it assesses the effectiveness of Random Forest and Multinomial Naive Bayes classifiers. According to the investigation, Random Forest surpasses Multinomial Naive Bayes in terms of accuracy and robustness across a variety of datasets and text volumes, whereas TF-IDF consistently exceeds BOW. The analysis also reveals that Reddit's social media data has the most predictive value. These findings emphasize how important data quality and advanced text representation are to enhancing stock price forecasting models.

1 INTRODUCTION

Since stock price forecasting affects investors, financial institutions, and market stability significantly, it is essential to financial research. For risk management, portfolio optimization and strategic investing to be successful, accurate projections are necessary. They help institutional and ordinary investors navigate the intricacies of today's financial markets. Accurate stock forecasts also improve capital allocation, balance supply and demand, and stimulate economic growth, all of which contribute to increased market efficiency. In today's fast-paced financial environment, forecasting is no longer just a competitive advantage but rather a need.

Traditional forecasting methods rely on historical price data and fundamental financial indicators. While these methods provide valuable insights, they often need to capture intricate, non-linear patterns and rapid market changes. The rise of machine learning represents a significant advancement in forecasting (Sun et al., 2024). Machine learning offers sophisticated tools to analyze large and diverse datasets, uncovering complex patterns that traditional techniques may overlook (Whig et al., 2024).

With the use of cutting-edge techniques that increase prediction accuracy, machine learning has completely transformed the forecasting of stock prices. Large volumes of data are easily managed and analyzed using machine learning, in contrast to traditional models that mostly rely on historical data and basic indications. Time series modeling, regression analysis, and neural networks are among the crucial techniques for identifying connections that traditional approaches could overlook.

One important advancement in this sector is the incorporation of social media data into forecasting algorithms (Ferdus et al., 2024). Unstructured data representing market rumors, public opinion, and emerging patterns is produced via platforms such as Reddit, Twitter, and financial news forums (Choi et al., 2024). This real-time data offers insights that cannot be obtained from past price data alone. The attitudes and beliefs of investors can be revealed

[a] https://orcid.org/0009-0001-2876-9562

Li, B.
Enhancing Stock Price Forecasting with Social Media Text: A Comparative Study of Machine Learning Approaches.
DOI: 10.5220/0013225200004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 331-336
ISBN: 978-989-758-726-9

through social media sentiment analysis, which regularly affects short-term market swings. Though different kinds of textual content can produce distinct outcomes, the usefulness of this idea is still up for debate. Predictability may be greatly increased or decreased by certain texts.

One of the main tasks in machine learning is text vectorization, which is converting unstructured text into a numerical format. Popular techniques include the Bag of Words (BOW) model and the Term Frequency-Inverse Document Frequency (TF-IDF) combination. With the BOW model, word frequencies are counted by treating text as a set of distinct words. It often produces sparse feature vectors that may overlook subtleties in meaning, despite its effectiveness. In contrast, TF-IDF gives each phrase a weight based on how important it is inside a document in comparison to the entire corpus. By emphasizing key phrases and reducing the prominence of common, uninformative terms, it offers a more advanced representation.

Another difficulty is choosing the right machine learning model. The efficacy of a model depends on its capacity to handle the particularities of social text data, such as slang, informal language, and different levels of information. For problems involving text classification and prediction, models like Random Forests and Multinomial Naive Bayes (MultinomialNB) offer clear advantages. As different models may perform better with different text properties and circumstances, selecting the best model requires thorough experimentation and validation.

Social media data from RedditNews, Asea Brown Boveri Ltd. (ABB), Google LLC (GOOG), Apple Inc. (APPL), and Exxon Mobil Corporation (XOM) are used in this analysis. Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are two text vectorization techniques that are used in this paper. In the tests, stock prediction performance is evaluated across different text volumes using Random Forest and MultinomialNB models (AAPL5, for example, stands for five randomly picked Apple news items). Based on the data from RedditNews, the results indicate that Random Forest and TF-IDF perform better than BoW and MultinomialNB in general.

Further analysis of the data's textual characteristics in this research revealed that prediction accuracy is strongly influenced by the association between textual data and the stock market. Text from RedditNews, for example, that shows a strong association with the market typically produces more amazing accuracy.

2 METHODS

In order to forecast stock movements using social text data, this article uses machine learning algorithms and text processing methodology. With a focus on the Random Forest and Multinomial Naive Bayes classifiers as well as the Bag of Words and Term Frequency-Inverse Document Frequency (TF-IDF) representations for text data, this section provides a thorough review of the techniques used.

2.1 Random Forest

For problems involving regression and classification, Random Forest is an ensemble learning technique (Sun et al., 2024). During training, it builds a large number of decision trees. In terms of classification, it outputs the mode of the classes; in terms of regression, it outputs the mean prediction of each individual tree. By replacing the samples in the training dataset, the system generates multiple decision trees. Predictions are combined to produce the final output after each tree is trained on a distinct bootstrap sample. Splitting is done on a random subset of characteristics at each node in the tree. By doing so, the resilience of the model is increased and the correlation between individual trees is decreased. The random selection of training data and attributes used to build each tree promotes variation among the trees.

2.2 Multinomial Naïve Bayes

Using the Bayes theorem, Multinomial Naive Bayes calculates the posterior probability of a class given the feature vector (Terentyeva et al., 2024). The conditional probabilities of the features that are assigned to each class are multiplied to estimate the probability of each class. Features are independent when conditioned on the class, according to the "naive" assumption. When features like word frequencies in text documents represent counts, Multinomial Naive Bayes is the best option. The distribution of words within each class is modeled using the multinomial distribution.

2.3 Bag of Words

Text is converted into numerical features using the Bag of Words (BoW) technique, which is a crucial text representation technique. While keeping the frequency of each word, it ignores word order and syntax. BoW breaks the text up into discrete words, or tokens. A distinct index is given to every token in a vocabulary. Next, the text is segmented into discrete words or tokens. In a vocabulary, each token is given a distinct index. Despite its inability to retain the text's semantic content or word order, BoW's simple design makes it simple to use and computationally efficient.

2.4 Term Frequency-Inverse Document Frequency

The BoW model can be improved by using TF-IDF, which shows a word's importance in a document in relation to its frequency in all papers. Frequency Term (TF). This metric counts the number of times a word occurs in a document. The ratio of a word's total number of occurrences to its frequency of occurrences in a document is used to compute it. The frequency of inverse documents (IDF). Throughout the entire corpus, this measure estimates a word's significance. The logarithm of the total number of documents divided by the number of documents that contain the word is how it is calculated. Lower IDF ratings for words that are often used in documents indicate that these words have less discriminating power. TF-IDF Score. The product of TF and IDF is the TF-IDF score. It rises when a word appears more frequently in a document and falls when a word appears more frequently in all documents combined. This draws attention to words that are common in one document but uncommon in the entire corpus.

3 EXPERIMENTAL RESULTS AND ANALYSIS

This research explores how different text representations and classification techniques perform and work when used to predict stocks using data from social media. The analysis examines various text volumes for AAPL, contrasts the performance of TF-IDF and RandomForest classifiers, assesses the influence of BoW versus TF-IDF representations, and compares the results across various companies and data sources, such as AAPL, XOM, GOOG, ABB, and RedditNews.

3.1 Dataset Description

The dataset consists of historical news headlines and stock price data, as shown in Table 1.

Table 1: The overview of dataset.

	Data range	Label
RedditNews	2008-06-08 to 2016-07-01	1: Adj Close value rose or stayed the same 0: Adj Close value decreased
ABB1	2014-02-011 to 2015-12-25	
APPL1	2014-01-02 to 2015-12-09	
APPL5		
APPL8		
GOOG1	2014-01-02 to 2015-10-09	
XOM1	2014-01-02 to 2015-12-09	

3.2 Experimental Setting

Three measures are used in this research to evaluate our model for stock prediction using social text data: Precision (P), Recall (R), and the F1 Score (F). These measures evaluate how well the model forecasts changes in stock price in response to news headlines (Krasnodębska et al., 2024). The percentage of accurate positive predictions the model makes is measured by weighted precision. Predicting whether a stock price will climb or stay stable indicates how accurate the model is. The model's ability to recognize every true positive example is shown by its weighted recall. It displays the number of times the model accurately predicted when the stock price actually increased or remained unchanged. Precision and Recall are balanced by the Weighted F1 Score. The integration of both measures into a solitary score facilitates the assessment of overall performance, particularly in the context of unbalanced datasets.

The data used in this experiment is split 80/20 between a training set and a testing set. With this method, the model can be trained efficiently on a large subset of the data while its performance is assessed on a different, hidden subset. By ensuring that the assessment takes into account the model's performance on fresh, real-world data, this technique helps to clarify the model's applicability.

3.3 Analysis

3.3.1 Comparison of AAPL with Different Text Volumes

When analyzing AAPL (Apple Inc.) stock predictions using different volumes of textual data, Table 2 shows varying performance metrics for both RandomForest and MultinomialNB classifiers across different text volumes: AAPL1, AAPL5, and AAPL8.

AAPL1: The performance shows a slight improvement with TF-IDF, yielding a precision of 0.51 and an F1 score of 0.47.

AAPL5: For AAPL5, there is a notable decline in all metrics compared to AAPL1, which is unexpected.

AAPL8: In the case of AAPL8, the data falls roughly between APPL1 and APPL5, making it challenging to ascertain the relationship between text volume and predicted performance due to fluctuating results.

When examining the effect of varying text amounts (1, 5, and 8) on AAPL stock prediction, it becomes evident that the amount of text significantly influences the performance of prediction models. However, as the text quantity increases to AAPL5 and AAPL8, the results demonstrate an unusual drop compared to AAPL1. Notably, the prediction performance strongly declines when transitioning from AAPL1 to AAPL5, as the Table 2 shows.

Table 2: Comparison of AAPL with different text volumes.

			BOW	TF-IDF
AAPL5	Random Forest	P	0.22	0.39
		R	0.46	0.46
		F	0.29	0.31
	Multinomial NB	P	-	0.45
		R		0.45
		F		0.39
AAPL8	Random Forest	P	0.22	0.53
		R	0.47	0.49
		F	0.30	0.40
	Multinomial NB	P	-	0.40
		R		0.43
		F		0.34
AAPL1	Random Forest	P	0.48	0.51
		R	0.47	0.49
		F	0.42	0.47
	Multinomial NB	P	-	0.46
		R		0.45
		F		0.43

3.3.2 Comparison of Random Forest and MultinomialNB

As shown in Table 3, the performance of the Random Forest and MultinomialNB classifiers varies significantly across different datasets and text representations.

All datasets show persistent good performance with Random Forest, especially when combined with TF-IDF. This suggests that it is efficient at handling a variety of textual characteristics and identifying intricate patterns. Bag of Words achieves substantially lower precision than Random Forest with TF-IDF for all text amounts. This implies that Random Forest is better at utilizing the deeper feature representation that TF-IDF offers.

MultinomialNB displays a variety of strengths. According to the statistics, Random Forest may be able to better capture the intricacies of social text data than MultinomialNB. The dataset it is used with determines how well it performs; occasionally, TF-IDF yields very good results. Nevertheless, MultinomialNB has constraints in certain scenarios, particularly with relation to handling smaller text volumes or employing Bag of Words.

For a given dataset and text representation, Random Forest typically performs better and more reliably. However, MultinomialNB struggles with complex or nuanced data and performs well in some cases.

Table 3: Comparison of RandomForest and MultinomialNB.

			BOW	TF-IDF
XOM1	Random Forest	P	0.21	0.21
		R	0.45	0.42
		F	0.29	0.28
	Multinomial NB	P	-	0.48
		R		0.46
		F		0.36
ABB1	Random Forest	P	0.21	0.21
		R	0.45	0.42
		F	0.29	0.28
	Multinomial NB	P	-	0.48
		R		0.46
		F		0.36
GOOG1	Random Forest	P	0.44	0.48
		R	0.43	0.48
		F	0.41	0.8
	Multinomial NB	P	-	0.47
		R		0.47
		F		0.47
	Random Forest	P	0.87	0.88
		R	0.85	0.88
		F	0.84	0.88
	Multinomial NB	P	-	0.89
		R		0.85
		F		0.85

3.3.3 Comparison of BoW and TF-IDF

The contrast between TF-IDF and BoW highlights how feature extraction and text representation affect prediction accuracy.

Word presence and absence are recorded by BoW, but word importance and document-specific distinctiveness are not taken into account. This constraint may reduce its capacity to discern between pertinent and extraneous phrases when predicting stocks. Generally, BoW produces poorer precision and F1 scores than TF-IDF, as has been seen in several datasets. BoW performs poorly in stock prediction tests that call on contextual comprehension and term relevance recognition since it just uses word frequency to represent text.

According to Wan et al. (2024), TF-IDF provides more complex text representations by allocating weights according to the significance of words in a given document in relation to the corpus as a whole. By using this strategy, forecast accuracy is improved and significant phrases are identified. Metrics showing superior F1 scores, recall, and precision with TF-IDF over BoW are consistently higher across datasets. This demonstrates that TF-IDF produces better predictions by capturing more significant patterns in text data.

To sum up, in most situations, TF-IDF outperforms BoW in prediction performance because it provides a more illuminating text representation.

The results of this investigation show that there are a lot of chances for incorporating social media text into stock price forecasting, especially when using complex text representations like TF-IDF and cutting-edge machine learning methods like Random Forest. Compared to simpler approaches like BoW, TF-IDF improves feature representation by highlighting the significance of phrases across documents, improving prediction accuracy.

Higher text volumes, like AAPL5 and AAPL8, do not always improve performance and may even decrease accuracy, according to the research. Given that higher volume may bring noise, this shows that text data quality and relevancy are crucial.

When utilizing TF-IDF, Random Forest performs more robustly than Multinomial Naive Bayes in handling complicated text features, consistently outperforming it in a variety of circumstances. In contrast, there are drawbacks to Multinomial Naive Bayes, especially when dealing with simpler representations and smaller text volumes.

4 DISCUSSIONS

This section compares the performance of RedditNews, AAPL, XOM, GOOG, and ABB in terms of social media text predicting. RedditNews stands out for having the best predictive accuracy, with precision and F1-scores over 0.84 and 0.88, respectively. Reddit performs better than other platforms because of its user-driven platform and abundant content, which highlights the importance of social emotion in financial forecasts.

By comparison, AAPL performs well at first but metrics deteriorate as text volume rises, especially in AAPL5, which had the lowest F1 score, recall, and precision. This shows that quality data is more important than quantity when it comes to producing superior results. With a precision of about 0.21 and an F1 score of about 0.28, both companies exhibit low-performance metrics for XOM and ABB. The reason for their poor performance could be attributed to less educational writing and common industry traits that restrict the depth of public conversation.

Table 4: The top 10 (frequence) words of XOM1 and RedditNews.

XOM1	RedditNews
Xom	US
URL	China
USER	EU
Exxon	UK
Mobil	India
Oil	Country
Stock	ISIS
Cvx	Police
Stock	Brexit
Corp	World
Appl	Government
energy	Russia

GOOG outperforms both XOM and ABB, benefiting from the practical application of TF-IDF, which enhances its predictive accuracy. The correlation between text features and stock prices is moderate for AAPL but stronger for RedditNews, illustrating the differing impacts of industry volatility and social sentiment.

Word frequency analysis indicates that industry-specific terms significantly influence predictive capabilities. For example, according to Table 4, they show a sharp contrast. RedditNews has the highest Precision, Recall, F1-score, so the prediction effect is good. Among which the high-frequency words are mostly the names of countries and regions, related

To political and economic events. While XOM with the worst prediction effect, are more focused on a small range and the disclosed events are often difficult to affect the stock market.

Enhanced data quality and multimodal techniques should be the focus of future study (Poojitha et al., 2024). Deep learning and other advanced machine learning approaches could improve predictions, especially for datasets with lesser performance (Xu et al., 2024; Kang et al., 2024). To properly use social media data for financial forecasting, ongoing improvement is necessary.

5 CONCLUSIONS

Prediction accuracy is considerably increased when social media text is included into stock price forecasts. The use of TF-IDF and Random Forest proves to be the most effective. The above findings indicate that while increasing text volume does not consistently improve performance, sophisticated text representations and robust classifiers such as Random Forest result in more reliable predictions. Notably, Reddit's social media data provides considerable predictive value. This underscores the importance of data quality and relevance in forecasting models.

REFERENCES

Sun, Y., Mutalib, S., Omar, N., & Tian, L. (2024). A novel integrated approach for stock prediction based on modal decomposition technology and machine learning. IEEE Access.

Whig, P., Sharma, P., Bhatia, A. B., Nadikattu, R. R., & Bhatia, B. (2024). Machine Learning and its Role in Stock Market Prediction. *Deep Learning Tools for Predicting Stock Market Movements*, 271-297.

Ferdus, M. Z., Anjum, N., Nguyen, T. N., Jisan, A. H., & Raju, M. A. H. (2024). The Influence of Social Media on Stock Market: A Transformer-Based Stock Price Forecasting with External Factors. *Journal of Computer Science and Technology Studies*, *6*(1), 189-194.

Choi, M., Lee, H. J., Park, S. H., Jeon, S. W., & Cho, S. (2024). Stock price momentum modeling using social media data. *Expert Systems with Applications*, *237*, 121589.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, *237*, 121549.

Terentyeva, Y. (2024). Sentiment Analysis, InSet Lexicon, SentiStrength Lexicon, Naive Bayes, Multinomial Naive Bayes, TF-IDF, Machine Learning. *International Journal of Open Information Technologies*, *12*(7), 32-37.

Wan, Q., Xu, X., & Han, J. (2024). A dimensionality reduction method for large-scale group decision-making using TF-IDF feature similarity and information loss entropy. *Applied Soft Computing*, *150*, 111039.

Krasnodębska K, Goch W, Verstegen J A, et al. Advancing Precision, Recall, F-Score, and Jaccard Index: An Approach for Continuous Gridded Data[J]. Recall, F-Score, and Jaccard Index: An Approach for Continuous Gridded Data, 2024.

Poojitha K, Rao L V K, Kumar K T D, et al. Predicting Stock Price Movements in Volatile Markets: A Multi-Model Fusion Approach[C]//2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE, 2024: 1-6.

Xu Z, Yu G. A Time Series Forecasting Approach Based on Meta-Learning for Petroleum Production under Few-Shot Samples[J]. Energies, 2024, 17(8): 1947.

Kang M. Stock Price Prediction with Heavy-Tailed Distribution Time-Series Generation Based on WGAN-BiLSTM[J]. Computational Economics, 2024: 1-20.