MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 3, pp. 982–992 (2025)

Sentiment-driven approach to refine stock price prediction

Tarsi M.[1], Ezzahoui I.[1], Douzi S.[2], Marzak A.[1]

[1] *Laboratory of Information Technology and Modeling, Faculty of Sciences Ben M'sick, Hassan II University, Casablanca, Morocco*
[2] *Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco*

(Received 28 March 2025; Revised 30 September 2025; Accepted 1 October 2025)

Stock price values are known for their volatility due to multiple factors making their predictability a difficult task. As social media posts and news can be considered as one of the major factors in price change, we aim in this paper to predict the next-day stock price of 4 different companies, using both social media and financial datasets that range from September 30, 2021, to September 30, 2022, as inputs. The datasets go through a preprocessing pipeline that includes sentiment analysis methods, where tweets are classified by employing TextBlob and finetuned RoBERTa to extract new features. The best model produces a 93% R^2 score and an RMSE value of 1.35.

Keywords: *stock market; stock price; Deep Learning; sentiment analysis; textBlob; fine-tuned Bert; LSTM.*

2010 MSC: 68T50 **DOI:** 10.23939/mmc2025.03.982

1. Introduction

Recent technological advancements, especially in Artificial Intelligence (AI) and its subfields Machine Learning (ML), Artificial Neural Networks (ANN), Deep Learning (DL), and Natural Language Processing (NLP) have significantly transformed various facets of human existence, including social, educational, and financial domains. Media and social media platforms have been instrumental in these shifts, as individuals increasingly depend on digital tools for guidance, trend awareness, and opinion expression. The accessibility of the internet has significantly impacted individual habits.

The stock market is a domain where information is essential for decision-makers, and the analysis of data to manage risks, negotiate complexities, and minimize losses is crucial. This domain is very arduous owing to its intrinsic complexity. The emergence of new technology has prompted various studies to tackle these difficulties, however progress has been incremental. Sentiment analysis has been thoroughly examined across diverse situations. One study [1] utilized sentiment analysis to assess restaurant user happiness, whereas another [2] integrated sentiment analysis from the COVID-19 period with stock market analysis. Furthermore, the study in [3] investigated the relationship between stock prices and public sentiment.

This case study seeks to forecast the subsequent day's stock price with Linear regression (LR) using Ridge model, Random Forest (RF) model and LSTM model, integrating sentiment analysis characteristics obtained from two established methodologies: Text Blob and a fine-tuned BERT model [4]. A comparison analysis is performed to evaluate the influence of sentiment analysis methodologies on the precision of stock price prediction models. The paper is structured as follows: Section 2 covers the fundamental concepts used in this research, Section 3 provides a literature review, Section 4 outlines the data and preprocessing pipeline and methodology, Section 5 presents the results, Section 6 for the discussion and Section 7 concludes the study.

2. Fundamental concepts

Machine learning (ML), a branch of artificial intelligence, where different approaches and their algorithms are applied depending on the study case, like regression, classification or clustering where the

aim is to solve problems using set of datasets in which the models are trained to find pattern inside the dataset to help experts or reduce the risks or augment the performance of some tasks, therefore the data are preprocessed to be fed to the models. The types of learning can be supervised, unsupervised, semi-supervised or by reinforcement. The supervised learning is when the target values are present in the data and used to guide the training of the model [5]. Several algorithms are used to apply the machine learning like linear regression, decision trees, random forest. Also, each case requires its own type of evaluation, like R^2 score with regression and accuracy with classification.

Deep learning (DL), also a branch of artificial intelligence, is modeled after the architecture and operation of the neural networks in the human brain. The similarity to biological brain systems is the reason certain models are referred to as Artificial Neural Networks (ANNs). Numerous deep learning models are available, such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Large Language Models (LLMs). Each of these designs is tailored for distinct applications, including time series analysis and computer vision. (LLMs) and they each can be used in different use cases like time series and computer vision.

2.1. LSTM

Long Short-Term Memory (LSTM) is a deep learning architecture designed to address a primary issue of Recurrent Neural Networks (RNNs), namely the disappearing or expanding gradient problem [6, 7]. LSTM employs three gates forget, update, and output that govern the information flow during training, ensuring that only pertinent information is transmitted to the subsequent cell [8, 9]. LSTM is a highly effective method for time series analysis and has been extensively utilized in stock market prediction research. Furthermore, it has been utilized in numerous research for text emotion classification.

2.2. Linear regression – ridge

Linear regression (LR) is a supervised learning algorithms, where the inputs are multiplied into coefficient to find the line in which predicted output produce minimal error score when compared with the true values of the target. The ridge linear regression is a linear regression algorithm, where its used to stabilize the linear regression problem's solution like multicolinearity or overfitting as it apply regularization l2, therefore the values of coefficient of features with least relation to target are minimalized [10].

2.3. Random forest

Random forest (RF) is an ensemble learning algorithms, where several decision trees are used to train each on subset of the original dataset where they are prepared by bootstrapping method, then averaging the prediction in case of regression or majority voting in case of classification [10].

2.4. Sentiment analysis

Sentiment analysis, a branch of Natural Language Processing (NLP), concentrates on discerning the prevailing emotion or sentiment conveyed in textual inputs such as comments, postings, publications, or titles. Sentiment analysis can be conducted by diverse methodologies, encompassing rule-based techniques and deep learning models. Frequently utilized instruments for sentiment analysis comprise:

TextBlob: A Python library for natural language processing applications, including text manipulation and fundamental operations. It has a sentiment analysis feature that produces a polarity score between -1 and 1. A score approaching -1 signifies negative sentiment, whereas a score around 1 denotes positive sentiment [11].

Bert: A robust pre-trained language model recognized for its profound contextual comprehension, as it examines text bidirectionally from both the left and the right. BERT is also amenable to fine-tuning, a crucial attribute for language models due to the variability of context across different academic disciplines [12]. In our research, we employed a refined BERT model [4], as described in the original reference.

The model was refined with the RoBERTa-base architecture on 3 200 000 comments categorized as 'Bullish' or 'Bearish'. Numerous research has utilized BERT for sentiment analysis. One study [13]

utilized BERT on a dataset of tweets, whereas another [14] investigated investor sentiment through BERT to assess its influence on stock returns. Comparative research [15] assessed Text Blob and BERT using product reviews from Egypt's Amazon platform.

In this study, we use two different models to classify the tweets sentiments, TextBlob as lexical-based model, and finetuned RoBERTa as semantic model. The objective is to determine which model can effectively translate information from social media, creating new extracted features that contribute to enhancing the performance of the price prediction model.

3. Previous works

A multitude of studies has been undertaken on this enduring issue, employing many data types financial, social, and economic to create models that can forecast price or return fluctuations. The data encompasses many time periods, notably including the COVID-19 era.

Table 1. Previous works.

Ref.	Title	Data time period	Regression models	Features involved
[16]	Stock Price Predictions with LSTM Neural Networks and Twitter Sentiment	November 30, 2020 to January 31, 2021	LSTM	Historical data, Twitter data (API)
[17]	Predicting Stock Market Indicators Through Sentiment Analysis On Twitter	February 13, 2017 to March 26, 2017	LSTM, Linear Regression	Close price, Twitter data (API)
[18]	Are Twitter sentiments during COVID-19 pandemic a critical determinant to predict stock market movements? A machine learning approach	January to May 2020	LSTM, ARIMA, Linear Regression, ARIMA	Macroeconomic indicators, Twitter data, Healthcare data, Stock data
[19]	Predicting stock price using LSTM and Social Media dataset	January 01, 2015 to January 01, 2020	LSTM	Historical data, Twitter data
[20]	A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing	January 2, 2015 to June 28, 2019	Multivariate regression, DT, Bagging, Boosting, RF, ANN, SVM, LSTM	Historical data, Twitter data (API)

Table 1 presents the data periods, models used, and data types involved in previous studies. Notably, the LR, LSTM and ensemble models are frequently utilized due to its effectiveness in addressing time series problems. However, the selection of sentiment detection methods is often overlooked, with studies typically opting for the most accurate sentiment models available. In sentiment analysis, the context and field of study are crucial considerations. While certain algorithms may seem robust, conducting a comparative study is beneficial to identify the optimal choice, as some models may require fine-tuning for specific contexts, such as the stock market. This study focuses on comparing the prediction results of Linear Regression using Ridge, Random Forest and LSTM models using features extracted through different sentiment detection methods. The goal is to evaluate whether the methodology of sentiment analysis and the accuracy of sentiment models significantly affect prediction outcomes and the performance of stock price prediction models.

4. Materials and methods

4.1. Data preprocessing

This section will delineate the dataset utilized and the preparation procedures implemented (see Figure 1). The dataset comprises data on the 'top 25 most viewed stock tickers on Yahoo Finance from

September 30, 2021, to September 30, 2022', as per its description. We opted to concentrate on four stocks, 'TSLA', 'TSM', 'BA' and 'META'. Consequently, working on 4 datasets independently, each stock on its own, although the columns remained unchanged, the number of rows was modified by restricting the data to encompass only items pertaining to the tickers as summarized in Table 2.

Table 2. Dataset size.

Data sets	Size
TSLA	37422
TSM	11034
BA	399
META	2751

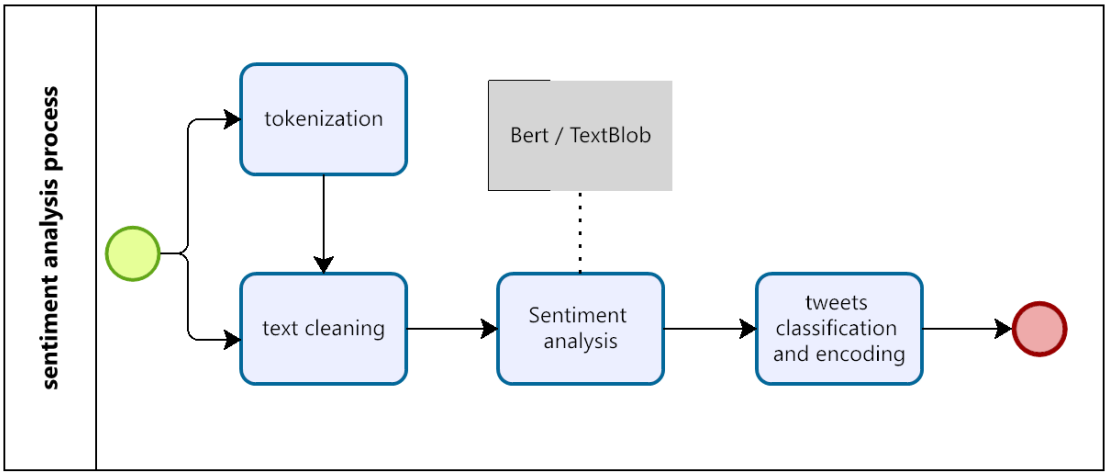

Fig. 1. Applying sentiment analysis to social media posts.

Data description. The data utilized in the present study and described in Table 3 was obtained from Kaggle [7]. The dataset comprises two files: the first provides tweet-related information, and the second encompasses historical stock data.

Table 3. Dataset contents.

Files	Features
stock_tweets.csv	Date, Stock Name, Company name, Tweet
stock_yfinance_data.csv	Date, Stock Name, Open, High, Low, Close, Volume, Adj Close

Data preprocessing steps. The models are trained on preprocessed data. Consequently, the data must be pristine, precise, and meticulously organized. Higher data quality enhances model training efficacy, perhaps resulting in superior performance.

Data preprocessing for sentiment analysis includes: Textual data cleansing which entailes eliminating special characters that may influence the analysis, Tokenization to facilitate subsequent preprocessing actions, Eliminating stop words like common terms which convey minimal substantive information, then apply sentiment analysis where Two methodologies are employed to classify tweets: Text Blob and a fine-tuned BERT model, which categorize sentiment as positive, negative, or neutral (for Text Blob).

Aggregating and merging data per date: The data is then aggregated daily and merged with financial features which produces missing values due to weekends where stock market is closed but there could be social features, in reverse some work days may not contains social features.

Addressing missing data: The values missing in the open and close features were replaced with last known values, as the open and close are the only financial features included in the training, however the values in the sentiment and social features were filled with 0 as no tweet apparently were present in that day.

Data split and scaling: for we split data into 3 sets training, validation and testing with 80/20 ration between training set and test set, then 80/20 ration again for training set and validation set. Another step was included is creating sequences for LSTM model, with 20 timesteps. After the data was split we applied the scaling using standard scaler.

4.2. Development of regression models for stock price forecasting

In this research, we constructed various models employing LR, RF, LSTM, for each algorithm we built 2 models, using features extracted from 2 different sentiment analysis methods. The initial model employs attributes obtained from the Text Blob polarity score to categorize tweets accordingly. The second model integrates attributes generated by a fine-tuned BERT model, with each tweet annotated with its emotion and score. The resultant features are encoded as independent variables, with each sentiment allocated a distinct feature (see Figure 2), thus, for all 4 datasets we have 36 models. For LR and RF models we used grid search for hyperparameters tuning, then we chose the best model to do the prediction.

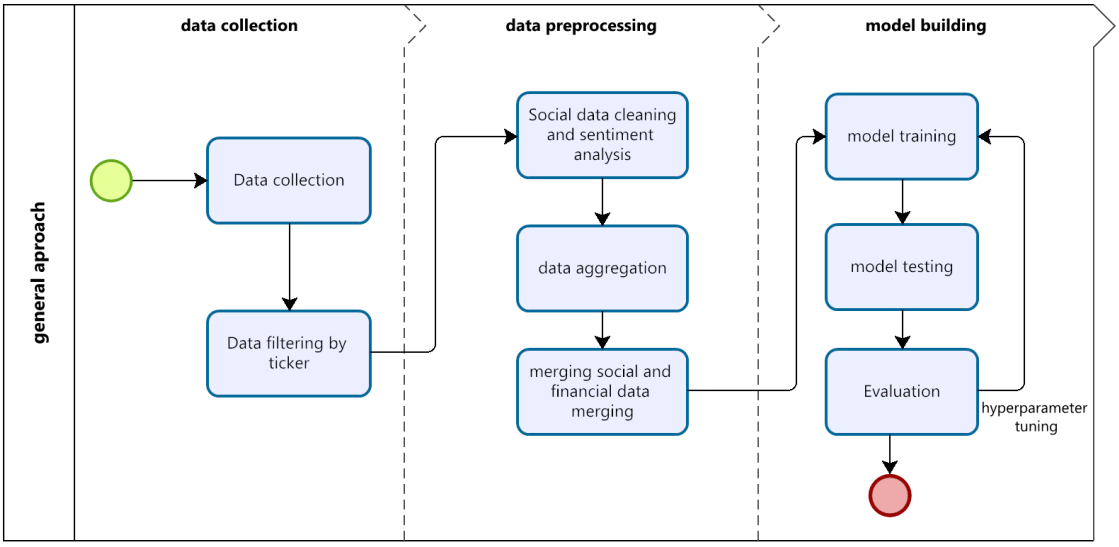

Fig. 2. The main workflow followed by our approach.

5. Results

Tables 4–7 summarize the results produced by LR, RF and LSTM models: one trained on sentimental features recovered using the fine-tuned BERT model and the other features obtained from Text Blob, while the last used financial features only.

For TSLA case Table 4, LR models showed the best result, an R^2 score of 74%, with textblob-based features and financial-only features and 73% with bert-based features while the RMSE value did not exceed 7.4 for all LR models. Random Forest produced an R^2 score of 67% with bert-based features, 69% with textblob-based features and 66% with financial-only features, however, the RF model without any social or sentiment features had an RMSE value of 8.33 exceeding other models that included social features. Finally, the LSTM model did not show good results in all case for all stocks, which can be explained by a need for more fine tuning and architecture improvements.

For TSM case (Table 5), LR models again showed the best result, an R^2 score of 93%, but this time with bert-based features and financial-only features while text-blob-based features produced 92%. The RMSE value did not exceed 1.4 for all LR models. The Random Forest models in TSM case did not produce good results as it did with TSLA where the R^2 score showed negative values, the same as LSTM models.

For BA case (Table 6), LR models showed the best result with an R^2 score of 89%, with textblob-based features and financial-only features while bert-based features produced 88%. The RMSE value did not exceed 3.9 for all LR models. However, the R^2 scores produced with Random Forest models in BA were poor compared to TSLA case, but better than TSM case, where values were between 27% and 35% with RMSE values that did not exceeds 9.8. The LSTM models again did not show good results.

Table 4. TSLA models' evaluation results.

Models	No social or sentiment features	Model with textBlob features	Model with finetuned Bert features
Linear Regression (Ridge)	R^2 score: 0.7491, MSE: 52.3429, RMSE: 7.2348	R^2 score: 0.7446, MSE: 53.2812, RMSE: 7.2994	R^2 score: 0.7375, MSE: 54.7685, RMSE: 7.4006
Random Forest	R^2 score: 0.6671, MSE: 69.4593, RMSE: 8.3342	R^2 score: 0.6973, MSE: 63.1596, RMSE: 7.9473	R^2 score: 0.6734, MSE: 68.1463, RMSE: 8.2551
LSTM	R^2 score: -0.5398, MSE: 220.8131, RMSE: 14.8598	R^2 score: -3.6691, MSE: 669.5879, RMSE: 25.8764	R^2 score: 0.0171, MSE: 140.9587, RMSE: 11.8726

Table 5. TSM models' evaluation results.

Models	No social or sentiment features	Model with textBlob features	Model with finetuned Bert features
Linear Regression (Ridge)	R^2 score: 0.9352, MSE: 1.8443, RMSE: 1.3581	R^2 score: 0.9254, MSE: 2.1220, RMSE: 1.4567	R^2 score: 0.9359, MSE: 1.8236, RMSE: 1.3504
Random Forest	R^2 score: -1.3558, MSE: 67.0083, RMSE: 8.1859	R^2 score: -1.3180, MSE: 65.9334, RMSE: 8.1199	R^2 score: -1.1607, MSE: 61.4584, RMSE: 7.8395
LSTM	R^2 score: -10.6887, MSE: 359.2834, RMSE: 18.9548	R^2 score: -2.5664, MSE: 109.6223, RMSE: 10.4701	R^2 score: -4.6820, MSE: 174.6515, RMSE: 13.2156

Table 6. BA models' evaluation results.

Models	No social or sentiment features	Model with textBlob features	Model with finetuned Bert features
Linear Regression (Ridge)	R^2 score: 0.8923, MSE: 14.3878, RMSE: 3.7931	R^2 score: 0.8925, MSE: 14.3616, RMSE: 3.7897	R^2 score: 0.8838, MSE: 15.5239, RMSE: 3.9400
Random Forest	R^2 score: 0.3531, MSE: 86.4207, RMSE: 9.2963	R^2 score: 0.3369, MSE: 88.5900, RMSE: 9.4122	R^2 score: 0.2739, MSE: 97.0058, RMSE: 9.8492
LSTM	R^2 score: -0.4247, MSE: 255.4082, RMSE: 15.9815	R^2 score: -8.7101, MSE: 1740.7531, RMSE: 41.7223	R^2 score: -6.7930, MSE: 1397.0761, RMSE: 37.3775

For META case (Table 7), LR models showed the best result with an R^2 score of 81%, with bert-based features and financial only features while textblob-based features produced 80%. The RMSE value did not exceed 5.2 for all LR models. However, the R^2 scores produced with Random Forest models in META case again showed very poor results. The same goes for LSTM models.

These results, derived from test data, indicate that LR models although simple but exhibit strong generalization which is encouraging for future applications. LR models exhibit strong generalization in all cases, however not the same can be said for RF models where it showed poor performance compared to LR. When models trained on bert-based features (TSM, META), RF shows negative values for R^2 score, which imply the possibility of the need of better feature engineering, as the loss values shows even smaller values in the LSTM when comparing the Model with fine-tuned Bert features and the models with no sentiment features (only-financial features: open and close), therefore adding sentiment features indeed add information to the model when training, but in need for more fine tuning and better architecture.

Figures 3 and 4 depict the anticipated price values in comparison to the actual pricing for models that included sentiment features in TSLA case. Figure 3 illustrates the model developed using features

Table 7. META models' evaluation results.

Models	No social or sentimentfeatures	Model with textBlob features	Model with finetuned Bert features
Linear Regression (Ridge)	R^2 score: 0.8192, MSE: 25.2694 RMSE: 5.0269	R^2 score: 0.8042, MSE: 27.3589, RMSE: 5.2306	R^2 score: 0.8163, MSE: 25.6712, RMSE: 5.0667
Random Forest	R^2 score: -4.7275 MSE: 800.4697, RMSE: 28.2926	R^2 score: -4.4740, MSE: 765.0358, RMSE: 27.6593	R^2 score: -4.9095, MSE: 825.9027, RMSE: 28.7385
LSTM	R^2 score: -17.5999, MSE: 2941.4907, RMSE: 54.2355	R^2 score: -25.0808, MSE: 4124.5719, RMSE: 64.2228	R^2 score: -16.0159, MSE: 2690.9952, RMSE: 51.8748

derived from BERT, whereas Figure 4 presents the findings from the model trained on features taken from Text Blob. In the comparison of the two models, the Text Blob-based model surpasses the BERT-based model in R^2 score, attaining 74% versus 73%. Nevertheless, a further examination of the plots uncovers significant disparities in the manner each model addresses particular price movements.

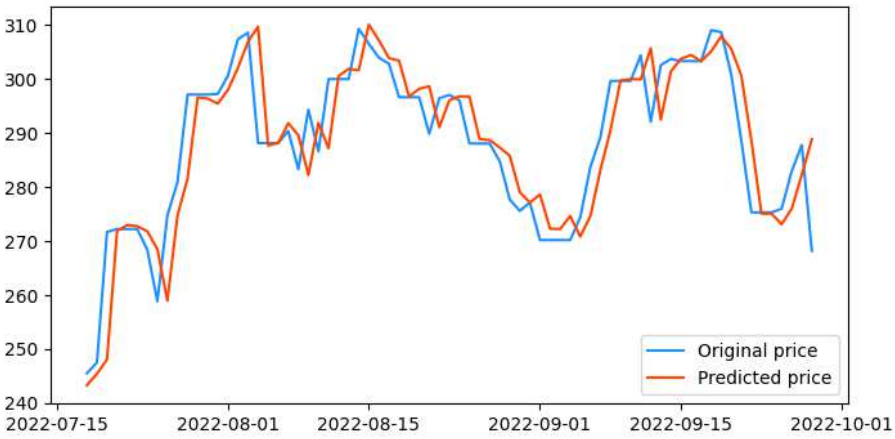

Fig. 3. Predicted price vs original price of the finetuned Bert-extracted-features based LR model, TSLA.

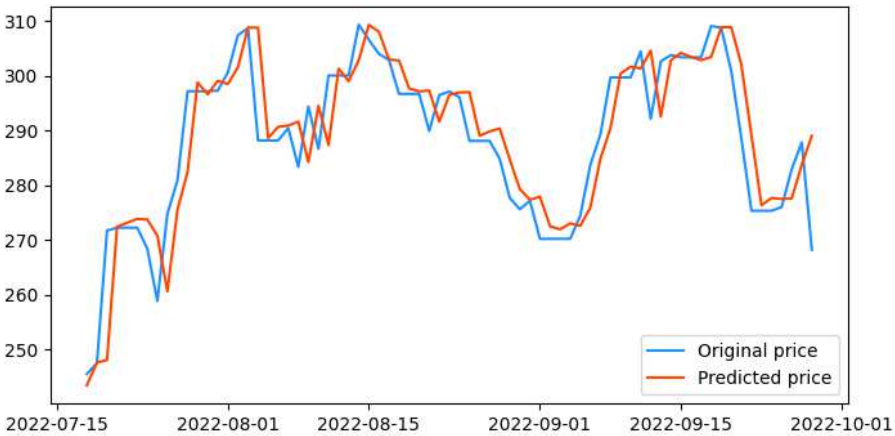

Fig. 4. Predicted price vs original price of the textBlob-extracted-features based LR model, TSLA.

Figure 5 and 6 depict the anticipated price values in comparison to the actual pricing for models that included sentiment features in META case. Figure 5 illustrates the model developed using features derived from BERT, whereas Figure 6 presents the findings from the model trained on features taken

from Text Blob. The opposite from TSLA case, the Bert-based model surpasses the textblob-based model in R^2 score, attaining 81% versus 80%. The same as TSLA case, a further examination of the plots uncovers significant disparities in the manner each model addresses particular price movements.

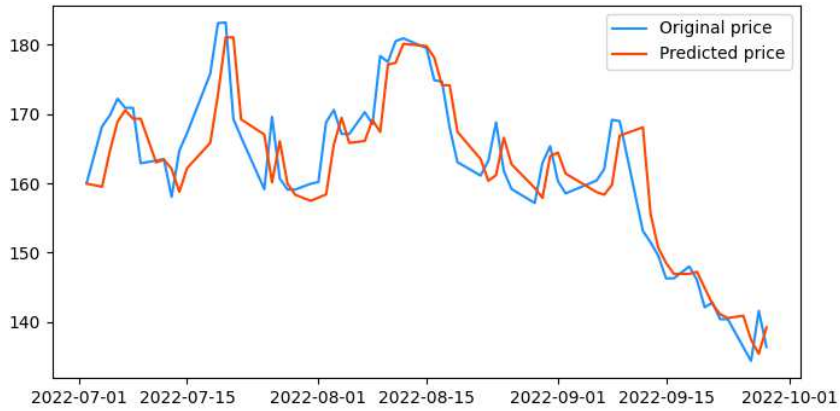

Fig. 5. Predicted price vs original price of the finetuned Bert-extracted-features based LR model, META.

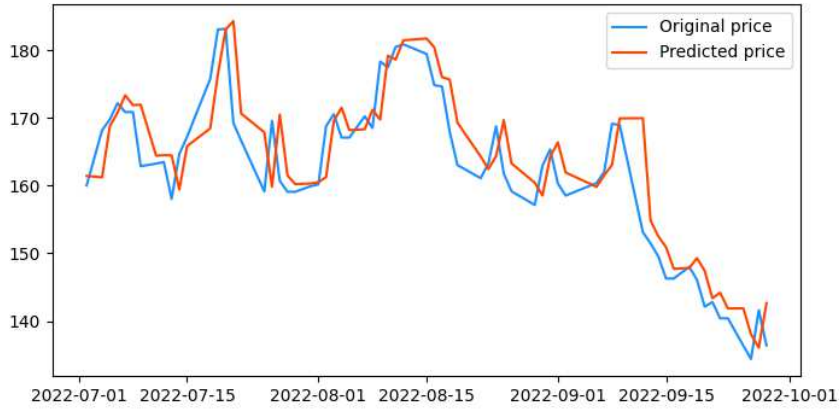

Fig. 6. Predicted price vs original price of the textBlob-extracted-features based LR model, META.

6. Discussion

Examining the scores produced by ALL Linear regression models of our approach as they showed the best performance among other algorithms used in this study case, the Text Blob-based model produces a better R^2 score when compared to the Bert-based model in case of TSLA and BA, but the opposite for TSM and META cases. However, the plots display more information on the progress of the prediction per data, when observed closely.

For TSLA case, in Figure 4, the Text Blob-based model closely aligns with the original price data during most intervals, potentially indicating greater accuracy at first observation. Nonetheless, it missed the low peak observed before 2022-08-01 by larger time, while showing some unfathomable fluctuations in the rising peak around same period. Furthermore, the model fails to capture the rising peak occurring around 2022-09-17. These inaccuracies can be detrimental, particularly during pivotal market transitions, despite the model's robust overall performance in other intervals. While the BERT-based model, illustrated in Figure 3, showed better performance in the periods mentioned compared to the textblob-based model, as the gap between prediction was lesser during the lower peak before 2022-08-01, the same for the peak around 2022-09-17, bert-based model could quickly catch up its original values, and the values did not exceeds the original values during the rising peak before 2022-09-15 compared to textblob-based mode potentially providing a more dependable forecast under specific market situations.

For META case, in Figure 6, the Text Blob-based model exceeds the values around the peak around 2022-07-15 and 2022-08-15 compared to the Figure 5, that illustrates the bert-based model. The same between periods 2022-09-13 and 2022-09-15, the Bert-based model showed smoother prediction catching the low trend and not exceeding the peak values although textblob-based model outperformed around 2022-09-01 catching up quickly with the rising trend.

This underscores that the selection of sentiment analysis techniques can profoundly influence the efficacy of stock price prediction models. Although all LR models demonstrate encouraging outcomes while RF models shows capability for improvement, their capacity to detect and react to subtle fluctuations during market peaks is a crucial aspect to evaluate. Both models necessitate additional fine-tuning, not only in the preprocessing of sentiment data but also in the optimization of the LSTM models' hyperparameters. These tweaks enable the models to attain greater predictive accuracy and effectively manage volatile market fluctuations.

As the results show, the models used to extract the sentiment features used in the price prediction produce different results which affect the forecasting and final model performance. One of the limitations is the scarcity of the social media labeled data of the stock market, for different periods of time, which mean a difficulty in creating suitable data for the models to train on, this can be backed with the appearance of different platform of social media where people of common interest can interact with each other creating probably new expressions that refers to certain events and this augment the complexity of the problem. Now lexical-based models or machine learning/deep learning-based models all start with data, and if we add sarcasm and irony connected to the industry itself, more complex models like finetuned Bert will be in need. Another case is the «memes» that introduces computer vision to the problem which lexical based models will not be able to catch, adding to the already challenging case problem.

7. Conclusion

This study utilized NLP processing on TSLA, TSM, BA and META (different industries) related tweets, extracting novel features through two sentiment analysis techniques: fine-tuned BERT and Text Blob. Subsequently, we consolidated the resultant sentiment components and integrated them with historical data. Following the resolution of missing values, normalization, and the generation of data sequences in case of LSTM, we constructed machine learning models LR and RF, a deep learning model with LSTM layers to forecast the subsequent day's stock price. The assessment outcomes from LR and FR models were encouraging, however more fine tuning is in need in case of LSTM as the model could not generalize very well.

The intricacies of the stock market are widely recognized, and in the current landscape, social media platforms offer significant data that can mitigate risks and furnish decision-makers with a more comprehensive view of market movements. This study illustrates that the selection of sentiment analysis techniques significantly impacts model efficacy. Although the Text Blob-based model exhibited superior overall accuracy, the BERT-based model showed a robust capacity to identify trends during significant peaks when incorporated into the LR model. Despite LR models demonstrating commendable performance during the review phase with test data, there exists potential for more enhancement regarding the deep learning approach model.

[1] Adi Laksono R., Sungkono K., Sarno R., Wahyuni C. Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naive Bayes. 2019 12th International Conference on Information & Communication Technology and System (ICTS). 49–54 (2019).

[2] Duan Y., Liu L., Wang Z. COVID-19 Sentiment and the Chinese Stock Market: Evidence from the Official News Media and Sina Weibo. Research in International Business and Finance. **58**, 101432 (2021).

[3] Pagolu V. S., Reddy K. N., Panda G., Majhi B. Sentiment analysis of Twitter data for predicting stock market movements. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). 1345–1350 (2016).

[4] zhayunduo/roberta-base-stocktwits-finetuned Hugging Face. https://huggingface.co/zhayunduo/roberta-base-stocktwits-finetuned (2024).

[5] Rebala G., Ravi A., Churiwala S. Machine Learning Definition and Basics. An Introduction to Machine Learning, G. Rebala, A. Ravi, and S. Churiwala, Eds., Cham: Springer International Publishing. 1–17 (2019).

[6] Hochreiter S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. **06** (02), 107–116 (1998).

[7] Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computation. **9** (8), 1735–1780 (1997).

[8] Tarsi M., Douzi S., Marzak A. Forecasting financial market dynamics: an in-depth analysis of social media data for predicting price movements in the next day. Social Network Analysis and Mining. **14** (1), 169 (2024).

[9] Aasi A., Imtiaz S. A., Qadeer H. A., Singarajah M., Kashef R. Stock Price Prediction Using a Multivariate Multistep LSTM: A Sentiment and Public En-gagement Analysis Model. 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). 1–8 (2021).

[10] Masini R. P., Medeiros M. C., Mendes E. F. Machine learning advances for time series forecasting. Journal of Economic Surveys. **37** (1), 76–111 (2023).

[11] Tutorial: Quickstart – TextBlob 0.18.0.post0 documentation. https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis.

[12] Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint arXiv:1810.04805 (2019).

[13] Bello A., Ng S.-C.,Leung M.-F. A BERT Framework to Sentiment Analysis of Tweets. Sensors. **23** (1), 506 (2023).

[14] Li M., Li W., Wang F., Jia X., Rui G. Applying BERT to analyze investor sentiment in stock market. Neural Computing and Applications. **33** (10), 4663–4676 (2021).

[15] Mahgoub A., et al. Sentiment Analysis: Amazon Electronics Reviews Using BERT and Textblob. 2022 20th International Conference on Language Engineering (ESOLEC). 6–10 (2022).

[16] Thormann M.-L., Farchmin J., Weisser C., Kruse R.-M., Säfken B., Silbersdorff A. Stock Price Predictions with LSTM Neural Networks and Twitter Sentiment. Statistics, Optimization & Information Computing. **9** (2), 2 (2021).

[17] Fuller A. Predicting Stock Market Indicators Through Sentiment Analysis on Twitter, report, University of Iowa. https://hal.science/hal-03516008 (2022).

[18] Jena P. R., Majhi R. Are Twitter sentiments during COVID-19 pandemic a critical determinant to predict stock market movements? A machine learning approach. Scientific African. **19**, e01480 (2023).

[19] Tarsi M., Douzi S., Marzak A. Predicting stock price using LSTM and Social Media dataset. 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). 1–4 (2023).

[20] Mehtab S., Sen J. A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing. TechRxiv. (2021).

Заснований на настроях підхід для уточнення прогнозування цін акцій

Тарсі М.[1], Еззахуї І.[1], Дузі С.[2], Марзак А.[1]

[1] *Лабораторія інформаційних технологій та моделювання, Факультет наук Бен Мсік, Університет Хасана II, Касабланка, Марокко*
[2] *Медичний та фармацевтичний факультет, Університет Мохаммеда V, Рабат, Марокко*

Значення цін на акції відомі своєю волатильністю через численні фактори, що ускладнюють їхню передбачуваність. Оскільки публікації та новини в соціальних мережах можна вважати одним з основних факторів зміни цін, у цій статті ми прагнемо передбачити ціну акцій 4 різних компаній на наступний день, використовуючи як вхідні дані набори даних із соціальних мереж, так і фінансові дані за період з 30 вересня 2021 року по 30 вересня 2022 року. Набори даних проходять конвеєр попередньої обробки, який включає методи аналізу настроїв, де твіти класифікуються за допомогою TextBlob та налаштованого RoBERTa для вилучення нових функцій. Найкраща модель дає показник R^2 93% та значення RMSE 1.35.

Ключові слова: *фондовий ринок; ціна акцій; глибоке навчання; аналіз настроїв; textBlob; налаштований Bert; LSTM.*