Stock Price Prediction Accuracy Utilizing Social Media Sentiment

Meldas Jaskelevičius

Vilniaus Gedimino technikos universitetas,
Saulėtekio al. 11, Vilnius, 10223, Lietuva
Meldas.jaskelevicius@gmail.com

Abstract. This paper examines the influence of social media sentiment from Twitter/X on Tesla (TSLA) stock price direction prediction. An automated pipeline was implemented using Apache Airflow, combining tweet scraping, FinBERT-based sentiment extraction, and a dual-task LSTM model performing simultaneous directional classification and magnitude regression. Over 278,000 hyperparameter configurations were tested. The best model achieved a weighted F1 score of 0.706 during training, yet live paper trading simulation over 63 trading days yielded only 46% accuracy. Post features were also explored and included in the prediction. The results suggest that raw daily sentiment aggregation from Twitter alone is insufficient for reliable stock price movement prediction.

Keywords: social media, sentiment analysis, stock prediction, LSTM, FinBERT, Twitter, machine learning.

1 Introduction

Predicting stock market trends remains a longstanding challenge due to the complex and unpredictable nature of financial markets [1]. While time-series analysis of historical market data has formed the backbone of quantitative investing, relying exclusively on this approach can create an echo chamber disconnected from real economic context [2]. The rise of social media has introduced a new dimension: millions of daily posts on platforms like Twitter reflect real-time investor sentiment that may complement traditional market signals.

This study investigates whether integrating Twitter sentiment data with historical price data improves the F1-score of a directional stock price classifier. Tesla (TSLA) was chosen as the target stock due to its high social media activity, volatility, and retail investor participation. The complete pipeline covers data collection, preprocessing, sentiment extraction using FinBERT, dual-task LSTM training with systematic hyperparameter optimization, and a paper trading simulation using Alpaca Markets API.

DOI: https://doi.org/10.15388/LMITT.2026.11

2 Related Work

A systematic review of 25 papers from 2023 to 2026 was conducted via Web of Science using the keyword combination "Stock Market AND Prediction AND Social Media AND Sentiment", which yielded 253 results from which 25 matched the research topic after manual filtering. Twitter/X is the dominant social media source, used in 13 papers.BERT type model is used in 12 papers for sentiment extraction and LSTM networks appear in 9 papers as the primary prediction model.

Reported F1-scores for LSTM-based approaches range from 0.447 to 0.85. Bacco et al. [3] found that Twitter-based sentiment does not uniformly improve predictions even with a combined LSTM-FinBERT framework. Liu et al. [4] showed that incorporating a neutral sentiment class with FinBERT and SVM improved the F1 from 0.536 to 0.653. A key finding across reviewed studies is that data quality and preprocessing strategies are often more influential than model architecture [5].

3 Methodology

3.1 Data Collection

Twitter posts containing "TSLA" keywords were scraped from September 1, 2022 to December 19, 2025 using Selenium with undetected-chromedriver. A total of 179,520 raw posts were collected, of which 162,380 remained after cleaning. A local-cloud hybrid architecture was used: Apache Airflow running in Docker detected missing date ranges and wrote JSON files; a local watcher script then executed the scraper and signalled completion back to Airflow. Historical OHLC price data was retrieved via the Alpaca Markets API.

3.2 Data Preprocessing

Text cleaning included emoji-to-text conversion, removal of @mentions and URLs, noise removal (punctuation), non-English character filtering, extra whitespace removal, and lowercase conversion. Each tweet was mapped to its next valid NYSE trading day: posts before 09:30 ET were assigned to the current trading day; posts during or after market hours were assigned to the next trading day. After cleaning, FinBERT (yiyanghkust/finbert-tone) was applied to assign a sentiment score of +1 (Positive), 0 (Neutral), or -1 (Negative) to each post.

Each record in the dataset represents a cleaned social media post enriched with sentiment scores and corresponding stock market data. The features included in the cleaned dataset are summarized in Table 1, which describes the variables used for subsequent analysis and model training.

Table 1. Features of the cleaned data set.

Adjusted Date	Date that has been adjusted based on New York time.
Checkmark	Displays if the user has a paid subscription for twitter or is recognized as an important person.
Cleaned At	When the posts were cleaned.
Cleaned Text	Cleaned up text.
Likes	The amount of likes a post has.
Replies	Shows the amount of comments made to the original post.
Reposts	Shows the amount of times this post was shared on twitter.
Sentiment Score	Score assigned to each post can be a -1,0,1.
Views	The amount of times a post has been seen.
ID	Identification of the cleaned post.
Original Tweet ID	Identification of the original pre cleaned tweet.
Timestamp	The date at which the post was made.
High	Highest Stock price of the adjusted date assigned to this post.
Low	Lowest stock price of the adjusted date assigned to this post.
Close	Closing stock price of the adjusted date assigned to this post.
Open	Opening stock price of the adjusted date assigned to this post.

3.3 Feature Aggregation

Individual posts were aggregated by adjusted trading date using the mean for sentiment score and engagement metrics (replies, reposts, likes, checkmark proportion), the first value for OHLC prices (identical across a trading day), and the count for post volume. This produced 828 daily observation vectors used as input sequences for the LSTM model. Table 2 summarizes the aggregated feature set.

Table 2. Features of the aggregated data set.

Adjusted Date	Date by which all associated posts were aggregated.
Checkmark Mean	The average of all checkmark scores.
Likes Mean	The average of all likes.
Replies Mean	The average of all Replies.
Reposts Mean	The average of all Reposts.

Sentiment Score Mean	The average of all sentiment scores.
Views Mean	The average of all views.
High First	First value of the high stock market price by date.
Low First	First value of the low stock market price by date.
Close First	First value of the close stock market price by date.
Open First	First value of the open stock market price by date.
Post Count	Amount of posts aggregated.

3.4 Model Architecture

A dual-task LSTM model was designed to simultaneously perform directional classification (price up or down) and regression (price change magnitude) for 3 future trading days. The shared LSTM backbone feeds two separate output heads: a sigmoid-activated classification head and a linear regression head. Combined loss is the sum of Binary Cross-Entropy (classification) and Mean Squared Error (regression) for each of the three forecast days. Xavier uniform weight initialization was used with a fixed random seed to ensure reproducibility.

3.5 Hyperparameter Optimization

Grid-based hyperparameter search was conducted across 278,211 configurations stored in a PostgreSQL database. Parameters tested are listed in table 3. Experiments were run under both random and fixed (non-random) weight initialization. An 80 / 20 percent train-test split was applied, 80% was used for training and the remaining 20% for testing. Model performance was ranked by a weighted F1-score: Day 1 weight 50%, Day 2 weight 33%, Day 3 weight 17%.

Table 3. Parameter explanation and the testing ranges.

Parameter	Explanation	Range
Hidden size	Indicates the amount of nodes within a layer.	1,2,3,4,5,6,7,8,9,10.
Number Of Layers	Show how many LSTM layers are used in the model.	1,2,3,4,6,7,8,9,10,12.
Learning Rate	Weight which affects the node weight calculation determining their change size each epoch.	0.1,0.9,0.01,0.001,0.0001.
Drop Out	A number that determines the amount of randomly disabled nodes within a layer.	0.01,0.05,0.1,0.3,0.5,0.9.

Parameter	Explanation	Range
Epochs	Determines how many iterations of training the model has to go through.	10,30,50,70,90,100,110,130,150,170,190,200,210,230,250,270,290,300.
Sequence Length	The length of the input data, meaning how many days in a row should the data set be divided into and used for training.	10,20,30,40,50,60,70,90,100,110,130,150,170,190,200,230,260,300.
Target	A simple variable which selects the target for predictions. In this experiment the target was set to Low.	Low,High,Open,Close.
Scaler	Sets the scaler to scale, of reposts, likes and replies, to give different importance to these metrics.	Robust, Minmax, Standard.
Bidirectional	Configures if bidirectionality should be on or off.	True or False.
Optimizer	Picks a different optimizer to determine the best one for accurate predictions.	Adam, Adamw, SGD.
Rerun	A variable only for random weighted nodes to determine the amount of reruns of the training process.	1 or 20.

4 Experiments and Results

4.1 Hyperparameter Analysis

Under random weight initialization, all hyperparameters produced near-identical statistics (mean F1 ≈ 0.36, std ≈ 0.21), indicating that weight randomness masked any real parameter effects. Under fixed initialization, clear patterns emerged. Number of LSTM layers showed the strongest signal: performance peaked at 4 layers (mean F1 = 0.53, std = 0.13) and declined consistently beyond that. Bidirectionality was the second most impactful parameter, raising the mean F1 from 0.22 to 0.43. Hidden size 1 was both the best-performing and most computationally efficient option. Dropout above 0.1 and epochs beyond 200 both degraded performance under fixed initialization. Optimizer, scaler, and learning rate had no measurable impact across any tested condition.

4.2 Sentiment Analysis

After Sentiment is assigned to posts using FinBert it can be seen that the sentiment distribution is heavily skewed toward the neutral class, a majority

of posts are labelled as neutral meaning the model could not properly determine the correct sentiment, that could be caused because of the text cleaning measures, that would mean that the model relies too much on emojis, punctuation marks or is configured to be too safe. These results can be seen in Fig. 1.

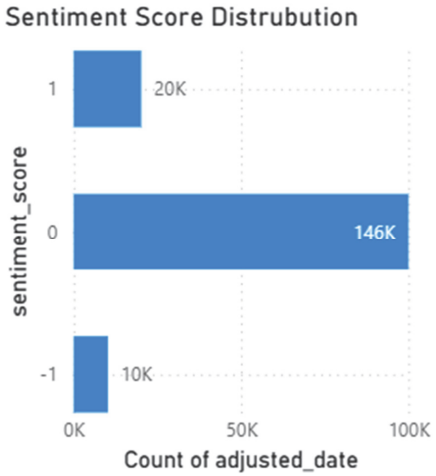

Fig. 1. Sentiment score distribution by the amount of tweets.

4.3 Feature Importance

Using Orange ranking node, feature importance to predicting the low price can be determined. This importance is measured using Universal Regression and Relief-F Algorithm. To simplify the higher the result the better, RReliefF can also have a negative value implying a negative correlation.

Table 4. Feature impact on predictive importance.

Feature	Univariate Regression	RReliefF
Adjusted Date	910.15	0.08
Checkmark	134.8	0.06
Sentiment score	150.97	0.06
Replies	133.87	0.05
Likes	188.87	0.04
Post count	0.08	0.04
Reposts	90.73	0.04

Taking a look at the results the biggest correlation is from the adjusted date implying strong time correlation meaning the data has a strong time series element. Universal regression shows that likes have a meaningful impact to the stock price meaning posts with high Likes influence investors. Thirdly Checkmark and Sentiment score do impact the prediction, this is a notable finding considering Checkmarks only show that the user has paid for twitter services and sentiment score should be higher, but the truth is it doesn't cause that big on an impact, a theory is that the post can be positive or negative , but reaction to it makes the difference, if a post is made by a checkmark user it naturally gets more exposure and if it gets a lot of replies or likes, it only then impacts stock price. Reaction to the post can also be negative or supportive.

To evaluate general feature behaviour over the full study period, the temporal trends of aggregated sentiment and engagement metrics were examined. Across the full dataset, no consistent visual relationship between individual engagement features and stock price movement was observed at a global scale. However, large-scale visualization across the entire period proved insufficient for identifying localized behaviour patterns, motivating further analysis using shorter time windows.

Looking at a graph of the 2 months the predictions focused on, Fig. 2 shows a clearer image of what is happening with features, it can be seen that periodically post counts spike drastically, but these spikes rarely coincide with checkmark users, meaning that more casual users or bots posted on those days.

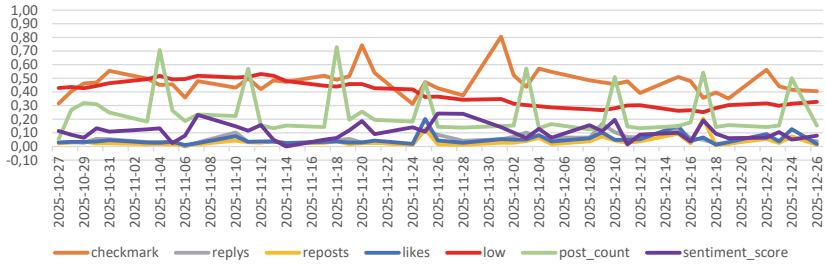

Fig. 2. Graph of minmax scaled features of the 2 month main predictive period.

4.4 Trading Results

The model was evaluated over 63 trading days within the research period. To assess directional accuracy, each prediction was compared against the actual price movement of Tesla stock - if the model predicted an increase and the price fell, the prediction was marked incorrect regardless of the action taken.

The model predicted upwards in 58 out of 63 cases, only predicting price movement downwards 5 times. Of the 27 correct upward predictions, the model genuinely captured rising price movements, however the 31 false positives cases where upwards movement was predicted but the market fell that represent the dominant failure mode. Overall, 29 predictions were correct and 34 were incorrect.

This gives an overall directional accuracy of 46%, which is below the 50% baseline of random chance. It is worth addressing the apparent contradiction with the F1 score of approximately 0.70 reported during hyperparameter optimization - the key difference is that the F1 score was computed during training and validation on the full dataset under controlled conditions, whereas these 63 trading results represent live inference on unseen data. F1 score also accounts for class imbalance differently than raw accuracy, and since the model heavily favours predicting “up”, it can achieve a high F1 on a dataset where upward movements are more frequent, while still performing poorly in practice.

Trading decisions are based on Day 1 predictions. Before placing any order, the system checks that prediction confidence exceeds 0.2 and the predicted price movement exceeds 0.1. If either threshold is not met, the system holds. Additionally, the classification and regression outputs must agree, an upward prediction must be paired with a positive magnitude, and a downward prediction with a negative magnitude. Conflicting outputs also result in a hold. When all conditions are satisfied, a buy or sell order is placed accordingly.

Regarding the action distribution, the model overwhelmingly chose to hold, signalling a buy only 14 times and a sell just twice across the entire period. This conservative behaviour is consistent with the known long-term upward trend of Tesla stock - a model trained on this data would naturally learn that holding or buying is more often correct than selling.

Breaking down correctness by action type in Table 8 makes the performance picture clearer. Buy signals were correct in 7 out of 14 cases

(50%), sell signals in 1 out of 2 (50%), and hold decisions in 21 out of 47 (45%). Every action category sits at or below coin-flip accuracy. This confirms that the model's trading signals carry no predictive edge - being correct half the time on buy and sell decisions, and below half on holds, is statistically indistinguishable from random guessing.

5 Conclusion

This study implemented an end-to-end automated pipeline combining Twitter sentiment data with historical TSLA price data to predict stock price direction using a dual-task LSTM architecture. Over 278,000 hyperparameter configurations were tested, with the best model achieving a weighted F1 score of 0.706 during training. However, live paper trading simulation over 63 trading days yielded only 46% directional accuracy, marginally below random chance, consistent with findings reported by Bacco et al. [3] and Ferraro & Sperlì [7], who similarly found Twitter-based sentiment signals to be inconsistent predictors.

The primary limitation was the strong upward prediction bias: the model predicted "up" in 58 of 63 cases, suggesting it learned TSLA's long-term bullish trend rather than actionable sentiment-driven signals. This aligns with the feature importance analysis, which identified adjusted date as the dominant predictive feature, confirming that temporal autocorrelation dominated over sentiment.

The FinBERT sentiment distribution, where the majority of posts were classified as neutral, likely diluted the predictive signal. As noted by Liu et al. [4], handling the neutral class explicitly is important for performance, and the text preprocessing pipeline (emoji removal, punctuation stripping) may have further reduced the sentiment information available to the model.

Hyperparameter analysis revealed that fixed weight initialization is essential for identifying meaningful parameter effects, under random initialization, all configurations produced statistically indistinguishable results. Within fixed initialization, bidirectionality and number of layers (optimal at 4) were the only parameters with a consistent and reliable impact.

These findings suggest that raw daily sentiment aggregation from Twitter alone is insufficient for reliable stock price direction prediction. Future work should explore engagement-weighted sentiment scoring, intraday sentiment segmentation, and structured news data as a complement to social media signals, as supported by Kim et al. [5] and Zhang et al. [6].

References

[1] Ahmed, D., et al. (2022). Analysis and Prediction of Healthcare Sector Stock Price Using Machine Learning Techniques. International Journal of Information System Modeling and Design, 13(9).

[2] Wang, S., et al. (2023). ALERTA-Net: A Temporal Distance-Aware Recurrent Networks for Stock Movement and Volatility Prediction. Proceedings ASONAM 2023, 538-542.

[3] Bacco, L., et al. (2024). Investigating Stock Prediction Using LSTM Networks and Sentiment Analysis of Tweets Under High Uncertainty: A Case Study of North American and European Banks. IEEE Access, 12.

[4] Liu, J. X., Leu, J. S., and Holst, S. (2023). Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and ensemble SVM. PeerJ Computer Science, 9.

[5] Kim, J., Kim, H. S., and Choi, S. Y. (2023). Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM. Axioms, 12(9).

[6] Zhang, P., Harris, R. D. F., and Zheng, J. (2025). GNN-based social media sentiment analysis for stock market forecasting and trading. Expert Systems with Applications, 291.

[7] Ferraro, A., and Sperlì, G. (2024). How does user-generated content on Social Media affect stock predictions? A case study on GameStop. Online Social Networks and Media, 43–44.