FINANCIAL MARKET FORECASTING WITH SENTIMENT INDICATORS

Khushi Omar, Annant Sharma, Sanya Baweja | Vijayasherly V | School of Computer Science and Engineering

Introduction

India's financial market includes primary and secondary markets for trading stocks, bonds, and currencies, regulated by bodies like RBI, BSE, and NSE.Economic factors and fintech growth are fueling market expansion and increased retail investor participation.AI techniques like LSTM and Random Forest help forecast stock trends using sentiment analysis and historical data patterns.This project aims to empower traders with predictive insights, advancing skills in fintech-focused AI and market forecasting.

Motivation

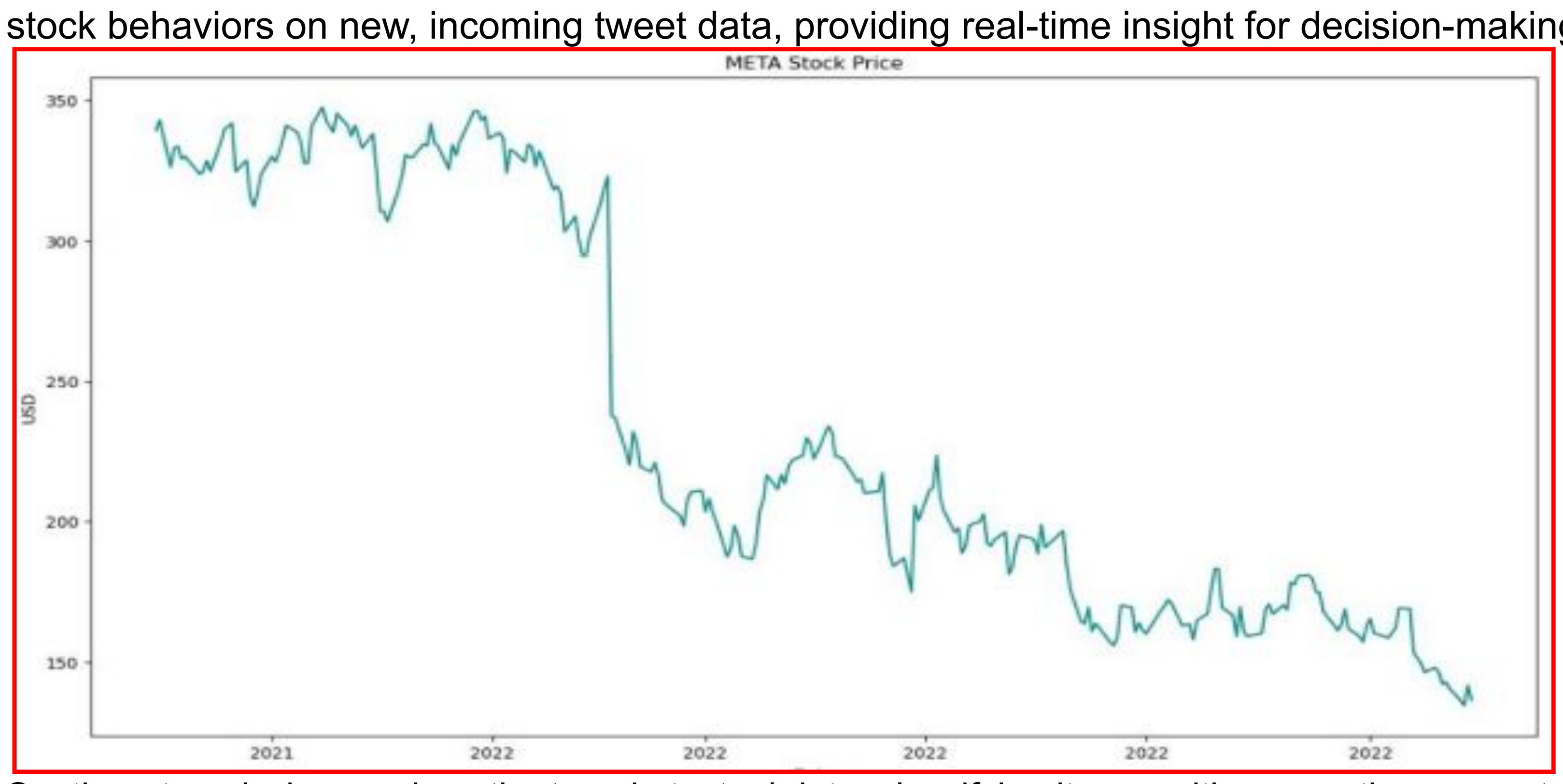
The stock market is driven by human emotions as much as data, with events like tweets or news affecting prices instantly. This project uses AI-powered sentiment analysis and time series forecasting to predict stock trends more accurately. Participants will gain hands-on experience in data science, finance, and AI tools like Python, LSTM, and NLP techniques. It fosters both technical and critical thinking skills, preparing individuals for careers in fintech, trading, and data science.

Scope of the Project

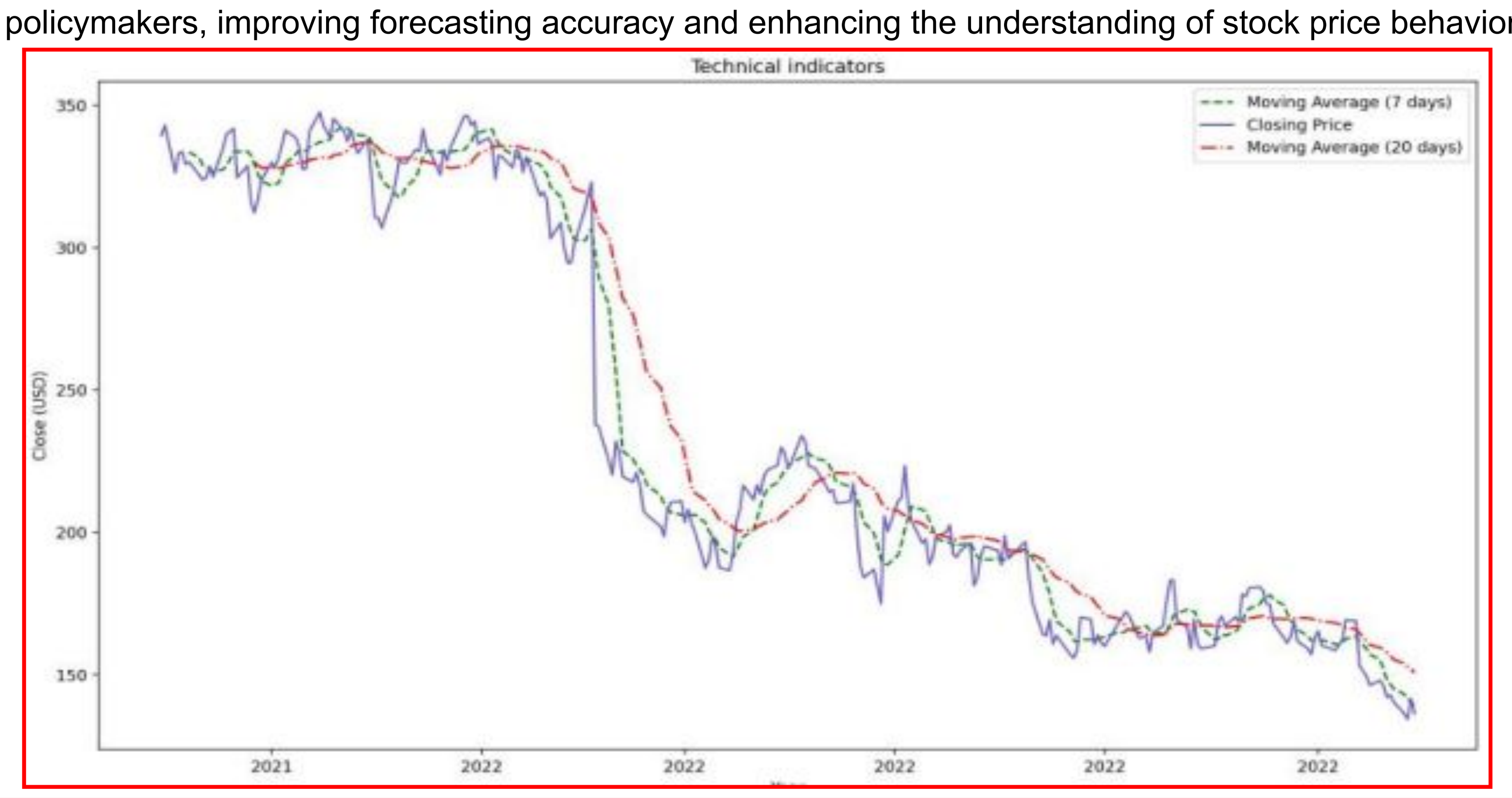
This project aims to enhance financial forecasting using AI, NLP, and sentiment analysis from platforms like Twitter for better market insights.Models like LSTM, Random Forest, ARIMA, and Prophet help predict trends across stocks, crypto, commodities, and currencies. It supports algorithmic trading with real-time sentiment triggers and helps detect fraud, promoting market stability. Built with tools like Python and TensorFlow, the system benefits retail traders, institutions, and regulators through data-driven transparency.

Methodology

- Input Data:The model utilizes a dataset of historical stock-related tweets, each containing the timestamp, tweet content, associated stock symbol (e.g., META, AMZN), and company name.
- Data Preprocessing:Raw tweet texts are cleaned by removing noise such as punctuation, special characters, and unnecessary whitespace, while sentiment scores are generated using the sentiment analysis tool.
- Feature Engineering:Key features such as sentiment polarity scores (positive, negative, neutral), compound scores, and timestamp-based attributes are extracted and combined with stock identifiers for each record.
- Kernel Selection:The Random Forest algorithm is selected as the model kernel due to its ability to handle non-linear data, reduce overfitting, and efficiently manage large feature sets via ensemble learning.
- Model Training:The Random Forest model is trained on the processed feature set to identify patterns and correlations between tweet sentiments and the stock's market implications.
- Hyperparameter Tuning:Model performance is enhanced by adjusting parameters such as the number of decision trees, tree depth, and feature sampling strategy through techniques like grid search and cross-validation.
- Model Evaluation:The trained model is evaluated using performance metrics including R-squared and Mean Squared Error to ensure accuracy and generalization on unseen data.
- Prediction on New Data:Once validated, the model is used to predict sentiment trends or related stock behaviors on new, incoming tweet data, providing real-time insight for decision-making.



Sentiment analysis examines the tone in textual data, classifying it as positive, negative, or neutral, using sources like financial news, social media, and corporate reports to capture public opinion and investor sentiment. Positive sentiment about a company's performance may indicate a price rise, while negative sentiment could signal a decline. Time series analysis, on the other hand, focuses on historical stock price data to identify trends and fluctuations, uncovering factors influencing market behavior over both short and long terms. Combining sentiment analysis with time series methods bridges the gap between psychological drivers and quantitative trends. The proposed model converts sentiment data into numerical scores and compares them with historical prices to evaluate predictive power, with time series methods contextualizing these scores within broader patterns. This hybrid approach enhances forecasting by linking investor sentiment with historical trends. Preliminary results show that combining sentiment and time series analysis improves predictions, as positive sentiment often aligns with price increases and negative sentiment correlates with declines, while time series analysis ensures historical data provides validation for sentiment-driven forecasts. Challenges include noisy data, misinformation, and the need for rapid processing, which can be addressed through robust preprocessing and reliable data sources. By integrating these methods, this study offers a valuable tool for investors and policymakers, improving forecasting accuracy and enhancing the understanding of stock price behavior.

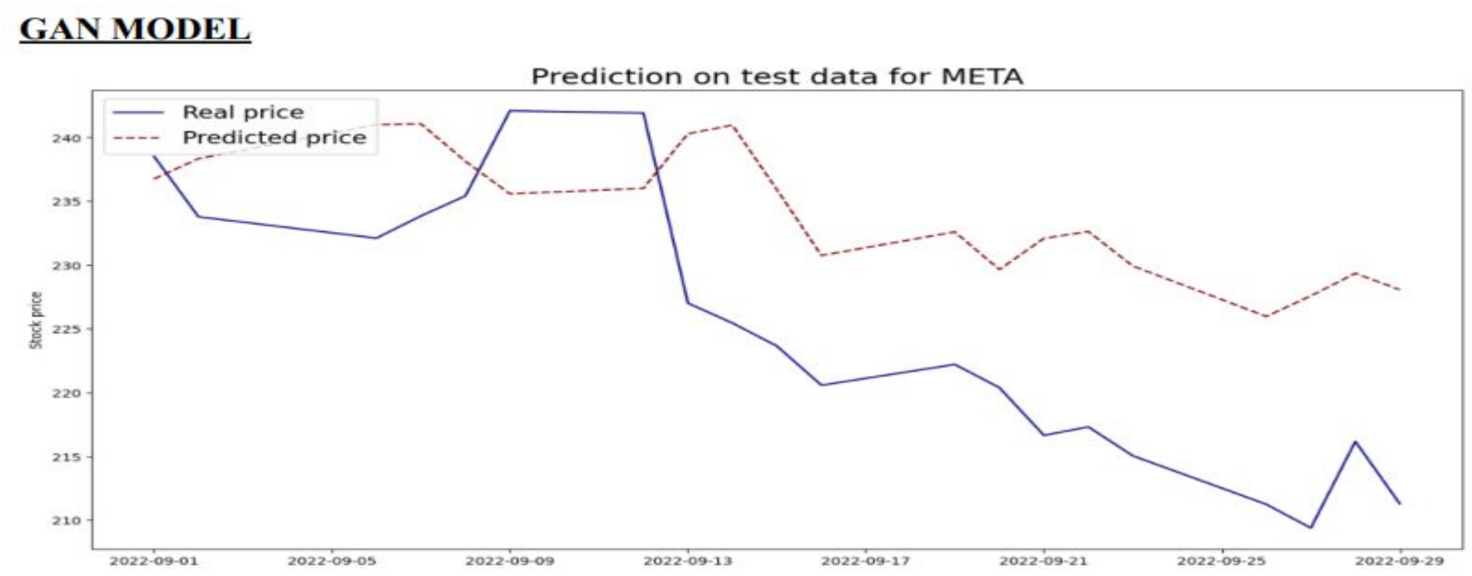


Results

Model Performance of META

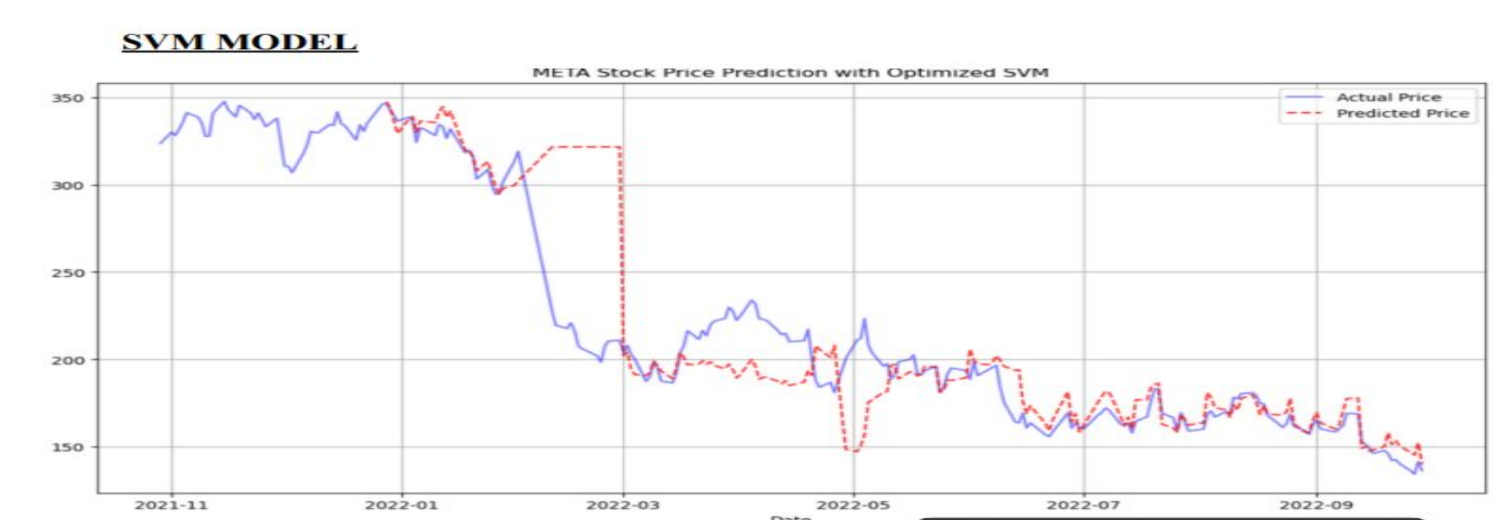
GAN MODEL

- The model effectively captures META's overall downward trend but shows a smoothing effect, missing sharp fluctuations.
- It tends to overestimate during sudden declines, indicating weak response to negative sentiment or news.
- Additionally, there's a noticeable prediction lag, with the model reacting slower than real-time market changes.



SVM MODEL

- The SVM model effectively captures META's long-term downward trend but struggles with sudden market changes, showing signs of lag and overfitting.
- It aligns moderately well with actual prices in stable periods but is less responsive to volatility.
- Predictions are often overly smooth or oscillatory, highlighting sensitivity to parameter tuning.



RF MODEL

- The model aligns well with small trend changes but struggles with high volatility, often under- or overestimating sharp spikes.
- It shows slight lag in predicting sudden market movements, indicating timing issues.
- Predictions are smoother compared to actual stock data, highlighting a trade-off between stability and responsiveness in time-series forecasting.

Conclusion

The project "Financial Market Forecasting with Sentiment Indicators" integrates sentiment analysis with machine learning and deep learning to improve market trend predictions. Models like SVM and Random Forest outperformed GANs, offering better accuracy, stability, and interpretability for sentiment-based forecasting. Sentiment data from Twitter enhanced traditional forecasting, enabling more informed decisions in volatile markets. While GANs are powerful for data generation, their high complexity and lower predictive accuracy made SVM and RF more suitable for this task.

References

[1]Machine learning in prediction of stock market indicators based on historical data from Twitter sentiment analysis(10 December 2013)

[11]https://www.kaggle.com/code/equinxx/stock-prediction-gan-twitter-sentiment-analysis/input?select=stock_tweets.csv

[12]https://www.kaggle.com/code/equinxx/stock-prediction-gan-twitter-sentiment-analysis/input?select=stock_yfinance_data.csv