Sentiment to Stocks: Forecasting Israel's Market Leaders

By Ahteshamuddin Mohammed, Soumya Shanigarapu and Teja Vineeth Reddy Yeramareddy.

INTRODUCTION:

In the dynamic landscape of global financial markets, recent geopolitical events, particularly the Israel-Palestine conflict, have instigated substantial fluctuations in stock prices, necessitating a nuanced analysis of investor sentiment. As financial stakeholders strive to comprehend the market dynamics amidst geopolitical uncertainties, this study endeavors to employ advanced machine learning methodologies, including sentiment analysis, to predict stock prices for five multinational corporations: Checkpoint (CHKP), Lundin Mining Corp (LUMI.TA), Mobileye (MBLY), NICE Ltd (NICE), and Poli Ltd (POLI.TA).

By amalgamating daily stock datasets from Yahoo Finance with sentiment data derived from Reddit discussions, our objective is to provide comprehensive predictions of stock price trends while shedding light on the public perception of these multinational corporations amidst geopolitical uncertainties. The parallels drawn from the success of machine learning models, including Random Forest Regression, Linear Regression, and Support Vector Regression, are integral to our approach, enabling the identification of the most accurate predictors of stock prices.

Beyond the immediate scope of the Israel-Palestine conflict, the versatility of our proposed pipeline renders it adaptable for sentiment and stock data analysis for any publicly traded company. As a valuable tool for future research and applications, the pipeline's adaptability is underscored by the potential integration of a user-friendly front end and parameterization of data pipeline components. Such enhancements promise heightened usability, efficiency, and scalability, offering tailored insights to users and seamless integration with other tools and technologies in the financial analytics domain.

PROBLEM DESCRIPTION:

In this project we seek to analyze the relationships between sentiment from Reddit discussions and stock prices of five key Israeli companies - Checkpoint, Lundin Mining, Mobileye, NICE Ltd, and Poli Ltd.

Our key goals are to:

1. Build an efficient data pipeline to extract relevant Reddit comments mentioning these companies over a defined timeframe.
2. Conduct sentiment analysis on the comments using TextBlob to categorize sentiment.
3. Integrate the sentiment data with stock price data from Yahoo Finance.
4. Compare machine learning models like Random Forest, Linear Regression and Support Vector Regression to predict stock price based on sentiment.

5. Ensure an adaptable and scalable framework that can incorporate other companies and social media platforms.

Overall, this project aims to provide an analytical solution to better comprehend the connections between geopolitical events, sentiment dynamics in social media, and Stock markets. The scalable data pipeline seeks to deliver actionable insights to empower stakeholders across evolving global scenarios.

Description of the Data (with Source):

For our project, we rely on two primary data sources: stock price data and Reddit comments data.

1) Stock Price Data:

Source: Yahoo Finance

Overview: Daily stock datasets for five Israeli companies are sourced from Yahoo Finance, covering a specified timeframe. The datasets provide crucial stock market information, including opening and closing prices, trading volume, and adjusted closing prices. These metrics serve as the foundation for exploring the connections between sentiment analysis outcomes derived from Reddit discussions and the trends observed in the stock prices of the selected Israeli multinational corporations.

2) Reddit Comments Data:

Source: Kaggle Dataset

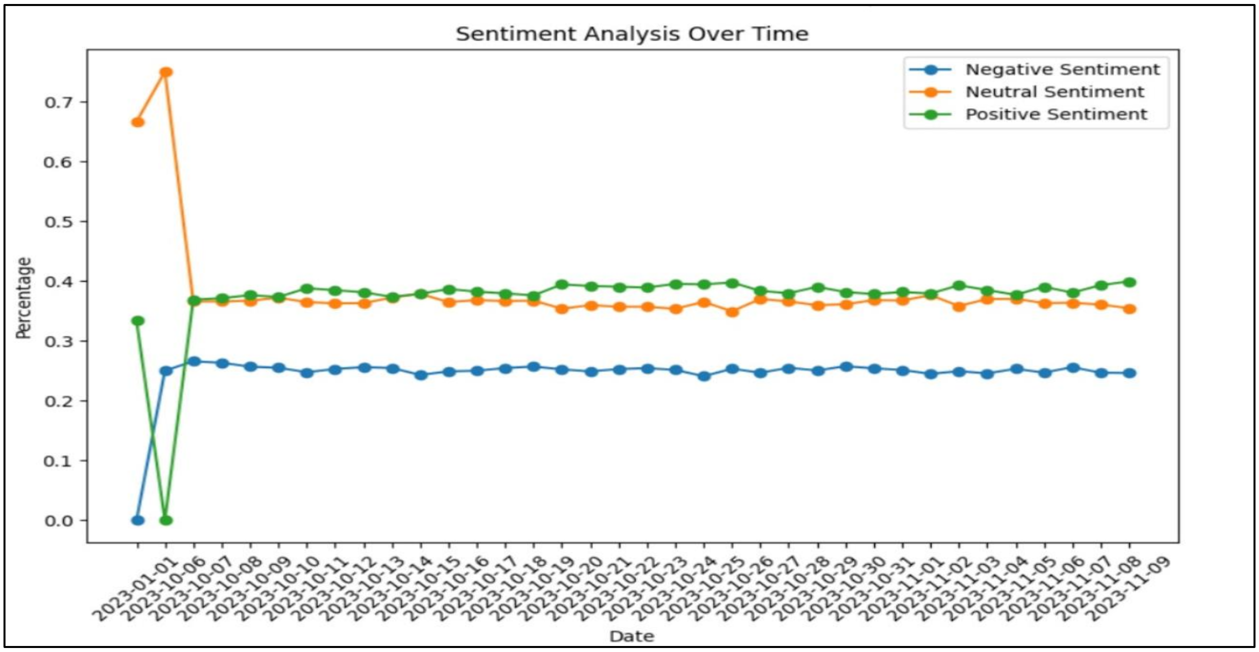
Overview: Our Reddit comments dataset is obtained from Kaggle, capturing daily conversations related to the Israel-Palestine conflict. Each entry includes a unique comment ID, user score, comment text, the online group, and a timestamp. This dataset allows us to track changing opinions over time in online discussions about the conflict. By applying sentiment analysis to these comments, we aim to uncover insights into user opinions and emotions surrounding the geopolitical situation.

Together, these datasets empower comprehensive analysis of the interconnections between sentiment dynamics extracted from social media discussions and corresponding stock price movements of leading Israeli corporations, within the politically tense climate surrounding the enduring conflict.

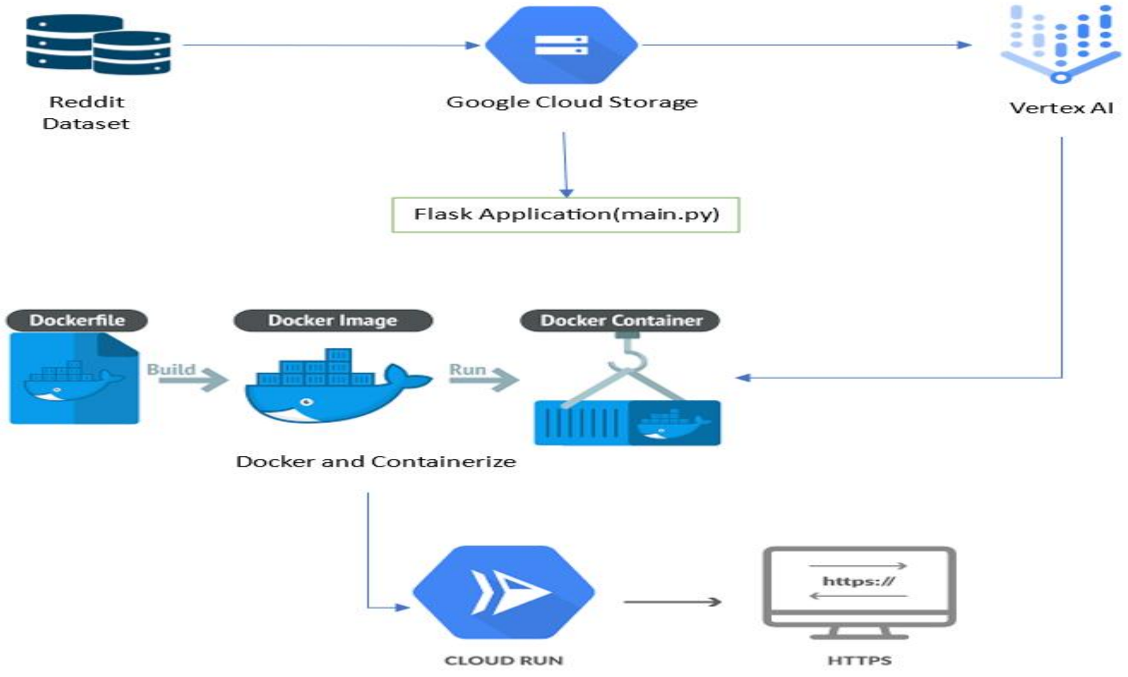
METHODOLOGY:

Sentiment analysis:

We utilize Text Blob's sentiment analysis capabilities to annotate Reddit comments as positive, negative, or neutral. By assigning these sentiment scores and leveraging timestamp data, we can systematically assess how sentiments within discussions related to the Israel-Palestine conflict change over time. This temporal sentiment analysis lays the groundwork for exploring potential correlations with stock price trends.



Application Flow:



Data Ingestion: We Acquired datasets from various sources and uploaded datasets to a designated Google Cloud Storage bucket.

Model Development and Training: We Utilized Google Vertex AI for model development and employed custom training methods for model training and these models are Trained using datasets stored in Google Cloud Storage. Then these trained models are stored in the Vertex AI Model Registry for future use.

Application Integration: We Developed a Flask application (main.py) as the core of the cloud-based solution and integrated the application with the Vertex AI Model Registry, leveraging Google Cloud libraries for seamless model retrieval.

Containerization: We employed Docker to containerize the Flask application, ensuring consistent and reproducible deployment environments.

Deployment: We Deployed the containerized application using Google Cloud Run which provides scalability and flexibility and Configured Cloud Run with the necessary credentials for secure interactions with other Google Cloud services.

Access and Testing: Accessed the deployed application through the generated URL, enabling testing and validation in a real-world environment.

Machine learning models:

In our pursuit of enhancing stock price prediction based on sentiment analysis, we implemented and compared three distinct machine learning models: Random Forest Regression, Linear Regression, and Support Vector Regression (SVR). These models were carefully chosen for their ability to decipher intricate relationships between sentiment features and stock prices.

Random Forest Regression: Leveraging ensemble learning, the Random Forest model excels in discerning complex and non-linear patterns within the data by amalgamating predictions from multiple decision trees.

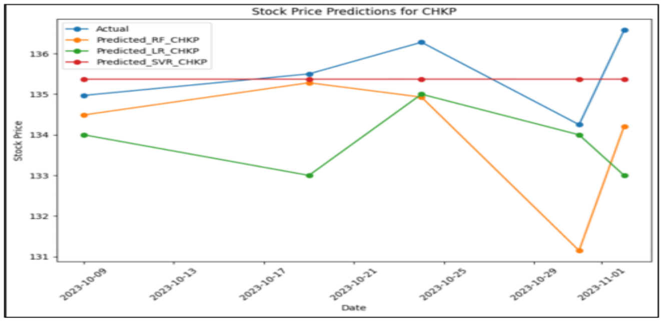
Linear Regression: Operating under the assumption of a linear relationship between features and stock prices, Linear Regression minimizes squared differences, providing a straightforward interpretation of feature impacts.

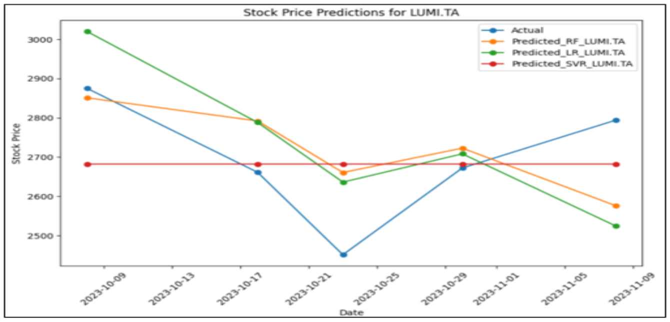
Support Vector Regression (SVR): Capitalizing on Support Vector Machines for regression, SVR navigates non-linear relationships by identifying hyperplanes within the feature space.

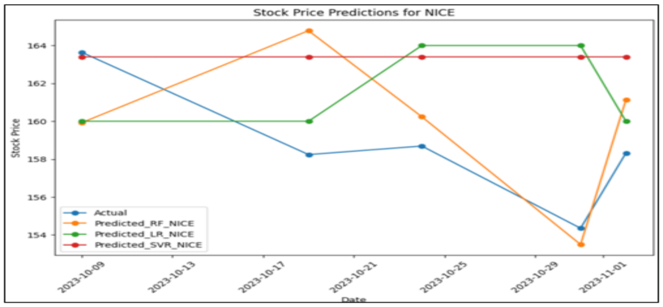
Our model evaluation exclusively employs the Mean Squared Error (MSE) as a robust metric. This metric serves as a benchmark to assess the predictive performance of each model. By scrutinizing the MSE outcomes, our goal is to pinpoint the most accurate model for predicting stock prices based on sentimental results.

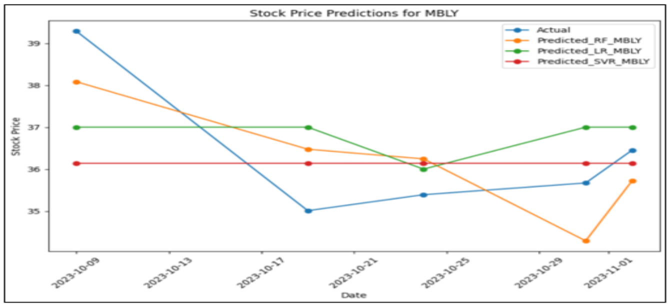
RESULTS:

Plots:









Evaluation results:

Evaluation Metric	Company	Linear Regression	Random Forest	Support Vector Regression
Mean Squared Error	CHKP	0.43	0.34	0.07
	LUMI.TA	0.29	0.22	0.20
	MBLY	0.23	0.13	0.24
	NICE	0.28	0.13	0.31
	POLI.TA	0.18	0.13	0.10

DISCUSSION:

The SVR model's performance was underwhelming, generating static predictions that failed to capture non-linear relationships in the data. This indicates limitations in adapting to the complex interconnections between sentiment features and stock prices.

Conversely, Random Forest Regression exceeded expectations. Its capability in modeling nonlinear patterns allowed superior predictive accuracy over SVR and Linear Regression. Resilience to noise and cross-dataset effectiveness cement Random Forest as a robust choice for sentiment-based stock price modeling.

While Linear Regression demonstrated good performance, Random Forest outmatched on overall results. This highlights the significance of selective modeling - preferring algorithms adept in unraveling the intricate dynamics of financial data.

Machine Learning Experiment Results Summary

Datasets: CHKP, LUMI.TA, MBLY, NICE, POLI.TA

Models: Random Forest, Linear Regression, Support Vector Regression (SVR)

Evaluation Metric: Mean Squared Error (MSE)

Key Findings:

In our exploration of machine learning models for stock price prediction based on sentiment analysis, a clear standout emerged – the Random Forest Regression model. Across all datasets and companies, Random Forest consistently outperformed both Linear Regression and Support Vector Regression (SVR). This emphasizes the effectiveness of Random Forest in handling intricate non-linear relationships and its resilience to noise.

Implications and Further Analysis:

While MSE offers a quantitative measure of error, it's advisable to delve deeper into model performance by exploring additional metrics such as R-squared or F1-score. Additionally, consider aspects like model interpretability and training time when making final decisions on model selection. This comprehensive analysis ensures a nuanced understanding of the models' strengths and limitations for informed decision-making in future applications.

REFERENCES:

1. Israel-related stocks fall globally as war escalates

- Source: Economic Times
- URL: [Israel-related stocks fall globally as war escalates - The Economic Times (indiatimes.com)]

2. Yahoo Finance

- Source: Yahoo Finance
- URL: [Yahoo | Mail, Weather, Search, Politics, News, Finance, Sports & Videos)

3. Reddit Data

- Source: Kaggle
- URL: [https://www.kaggle.com/datasets/asaniczka/reddit-on-israel-palestine-daily-updated/data]

4. TextBlob: Simplified Text Processing

- URL: [https://textblob.readthedocs.io/en/dev/]

5. Pandas Linear Regression

- Source: Scikit-learn Documentation
- URL: [https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html]

6. Pandas Support Vector Regression (SVR)

- Source: Scikit-learn Documentation
- URL: [https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html]

7. Pandas Random Forest Regressor

- Source: Scikit-learn Documentation
- URL: [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html]

8. Project files:

- URL: [https://drive.google.com/file/d/1oFeOucGwYYfgjQmP6LycBmginGfNNHcW/view?usp=sharing]

APPENDIX 1: Contributions

Team Member	Contribution
Soumya Shanigarapu	Data Collection, Data Processing & Data Ingestion & report
Ahteshamuddin Mohammed	Cloud Vertex AI model generation, ML models, Containerization of Application & report
Teja Vineeth Reddy Yeramareddy	Deployment, Flask application (main.py code part) & report.