Social Media based Stock Market Analysis using Big-Data Infrastructure

Vishakan Subramanian Venkataraman Nagarajan Shashanka Venkatesh Dr. Sujaudeen N

Department of CSE, Sri Sivasubramaniya Nadar College of Engineering
Final Year Project, May 2022

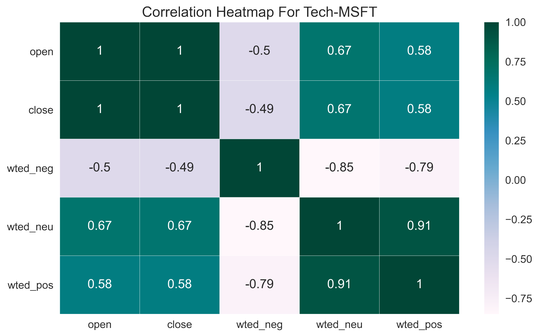
Highlights of Proposed Model

- State of the art BERT Model is used for sentiment analysis.
- Identifies exact correlation between industry trends from Twitter space and stock value.
- Predicts the next day closing value of a given stock.

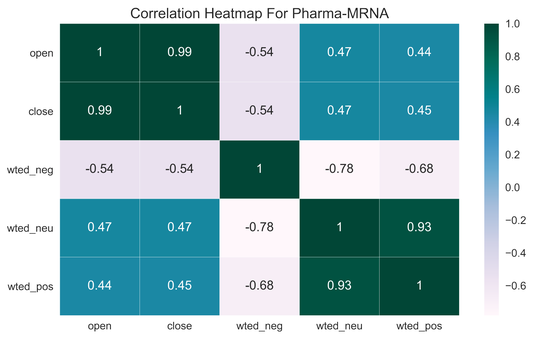
Challenges in assessing correlation:

- Tweets are not spread evenly for every sector.
- Missing Tweet data on certain dates & keywords.
- Stock market data unavailable on holidays & weekends.

Performance Metrics of Correlation Analysis



(a) Tech - (NASDAQ: MSFT)



(b) Pharma - (NASDAQ: MRNA)

Figure 1. Correlation Heatmap between Tweets and Stock Market.

Functional Modules and Dataset Description

- Data Collection & Pre-Processing
  - Collection of historical Tweet data
  - Collection of historical stock data
  - Performing data cleaning - i.e. removing duplicates, emoticons, links, etc. from Tweets.
  - Performing data imputation - approximate the stock market values for missing dates.
- Data Ingestion - Using Apache Kafka
- Data Processing - Using Apache Spark
  - Tweet Sentiment Analysis
  - Aggregation of date-wise social media sentiment
- Data Analysis
  - Correlation Analysis
  - Stock Market Prediction

- Four market sectors are taken: EVs, Tech, Oil & Pharma.
- Four candidate companies are considered: Tesla Inc., Microsoft Inc., Exxon Mobil Corp. & Moderna Inc.
- Tweet data was collected using Twitter-API with Academic Research level access.
- Stock market values - (Open, Low, Close, High) were collected for those companies using Polygon API.
- Period of study: Mar 2020 - May 2022

Sector	Data Points
Electric Vehicles (EVs)	116, 355
Oil	53, 253
Technology (Tech)	632, 181
Pharmaceuticals (Pharma)	313, 407

Table 1. Data Points Count

Proposed Model for Twitter Sentiment-Stock Analysis

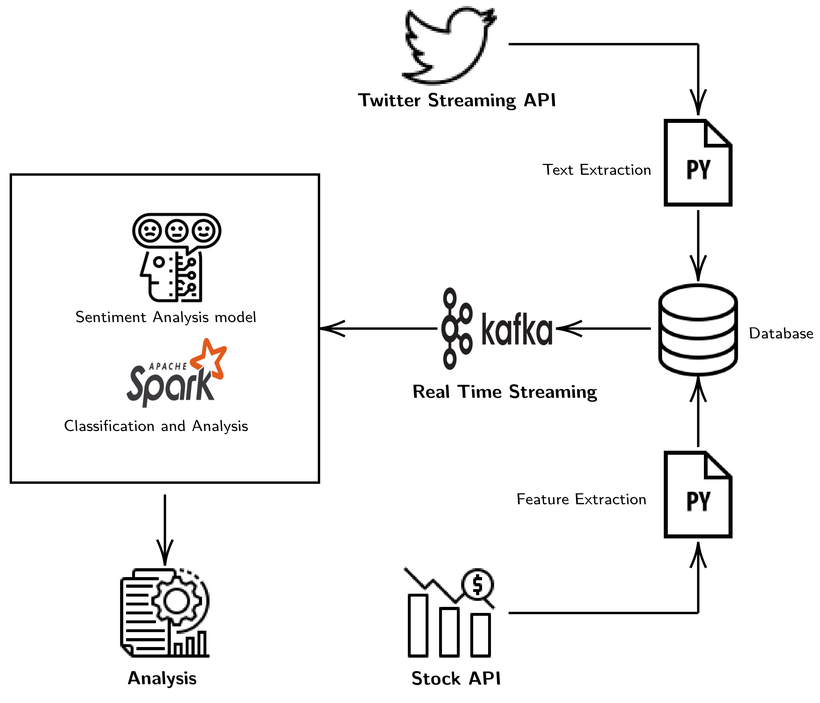


Figure 2. Proposed Architecture

Correlation Analysis and Prediction

- The collected Tweet data is cleaned, processed and segregated w.r.t. its industry sector and merged with the respective industry's representative stock value for each day.
- Then correlation analysis is performed between weighted sentiments and closing price of the stock.
- Gradient Boosting Algorithms (XGBoost and CatBoost) are used to predict the next day closing price of the stock.

XGBoost Prediction and Metrics

RMSE	0.0471
MSE	0.0022
MAE	0.0377
R2 Score	0.8381
Explained Variance Score	0.8427
Max Error	0.1403

Figure 3. Tech - (NASDAQ: MSFT)

CatBoost Prediction and Metrics

RMSE	0.0337
MSE	0.0011
MAE	0.0251
R2 Score	0.8461
Explained Variance Score	0.8642
Max Error	0.1491

Figure 4. Pharma - (NASDAQ: MRNA)

References

- Johan Bollen, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market". In: *Journal of Computational Science, Vol. 2, No. 1* (Mar. 2011), pp. 1–8
- Chungho Lee and Incheon Paik. "Stock market analysis from Twitter and news based on streaming big data infrastructure". In: *IEEE 8th International Conference on Awareness Science and Technology, iCAST 2017, Taichung, Taiwan, November 8-10, 2017*. IEEE, 2017, pp. 312–317
- Andreas Kanavos et al. "An Entropy-based Evaluation for Sentiment Analysis of Stock Market Prices using Twitter Data". In: *15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020, Zakynthos, Greece, October 29-30, 2020*. IEEE, 2020, pp. 1–7