Social Media based Stock Market Analysis using Big-Data Infrastructure

Group ID: G2 10

Shashanka Venkatesh
Venkataraman Nagarajan
Vishakan Subramanian

Mentor: Dr. N. Sujaudeen

SSN College of Engineering, Chennai

May 26, 2022

Outline

Motivation

- In recent years, news events are influencing stock values to a great extent.
- A greater influence can be felt from social media trends.
- Unexpected stock market fluctuations are becoming more frequent.
- Immediate reactions to the stock market can be seen when incidents are reported/propagated using extreme sentiments.

Problem Statement

- In a nutshell, we wish to analyse the correlation between social media trends & news articles relating to specific sectors and its effect on their current stock valuations.
- To develop a big-data architecture that is capable of handling voluminous information from sources like Stock APIs, Twitter etc.
- To assess and quantify the effect of social media on the economics of different sectors.

Justification

Observation

A lot of people invest in the stock market as a means to make money quickly. Shareholders react adversely and immediately towards their owned stock of relevant companies based on news trends, adding volatility (risk) to their valuation.

Conclusion

Presently, with the majority of the world being online, people tend to be more engaged on social media & use it a a source of news. Finding the correlation between such events & stock valuations is hence justified.

Literature Survey

- Lee et al.'s[5] research delves into analyzing Twitter data, by classifying it into categories and performing sentiment analysis to predict the trend of stock prices, and comparing it to reality, to find the correlation between the two.
- Scope of Further Improvement
  - Statistically insignificant volume of data, which reduces the confidence in their conclusion.
  - An increase of nearly 70 million users on Twitter between 2017 and 2021, giving reason to believe that the results may be different.
  - Usage of better machine learning techniques to improve classification & sentiment analysis.

Literature Survey (contd.)

- Kanavos et al.[4] have used the Stanford Core NLP library to perform text pre-processing along with Naive Bayes classifier to perform sentiment analysis on Twitter data for stock price predictions
- Scope of Further Improvement
  - Usage of better NLP tools - like BERT.
  - Usage of other cutting edge ML models for prediction.

Proposed System

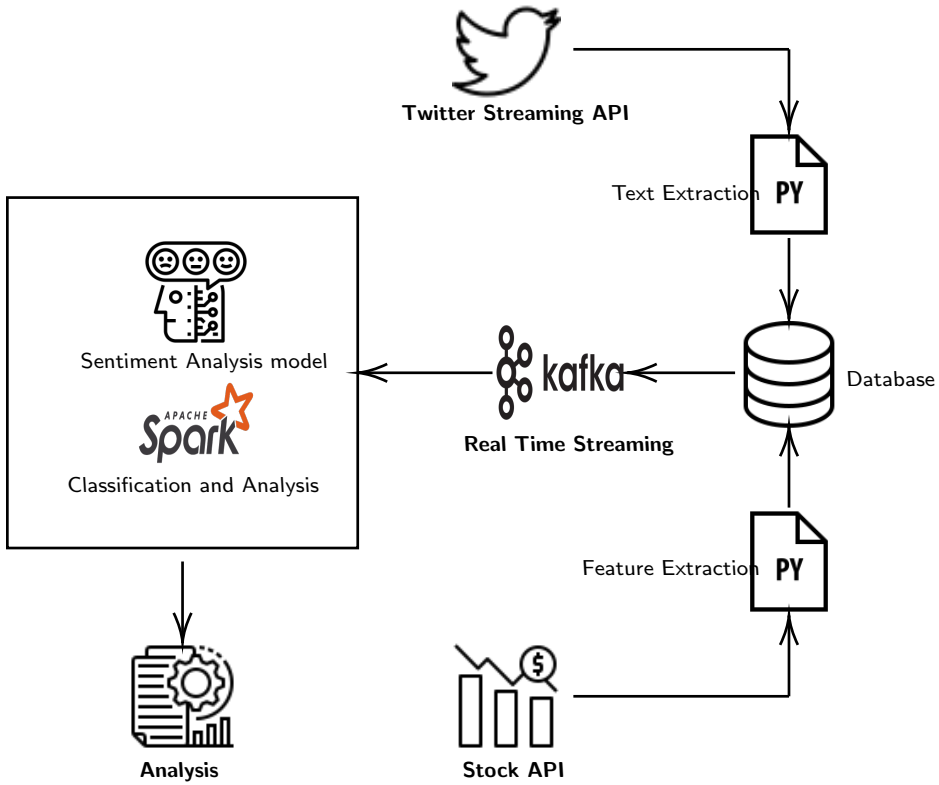


Figure: Proposed Architecture

Modules Involved

- 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

Techniques Used

- Tweet Data Collection: Official Twitter API, with Academic Research Access
- Stock Data Collection: Polygon API
- Data Storage: SQLite Database
- Data Ingestion: Apache Kafka
- Data Processing: Apache Spark (PySpark)
- Data Analysis and Viz.: Python

Techniques Used (contd.)

- Tweet Sentiment Analysis: Cardiff NLP's TimeLMs based RoBERTa Model.
  - The above model is pre trained on 124M Tweets collected between 2018 - 21, which is similar to our period of study. It has good results in the TweetEval task.
- Two cutting-edge boosted decision tree models: XGBoost [1]& CatBoost [12].
  - From Nayak et. al's [9] research, it can be seen that Boosted Decision Trees perform the best for this use-case when compared to other algorithms like Regression & SVM.

Understanding The Data

- 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 based on relevant hashtags like #MSFT, $TSLA etc.
- Stock market values - (Open, Low, Close, High) were collected for those companies.
- Period of study: Mar 2020 - May 2022

Table: Data Points Count

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

Understanding Processed Data

- Collected Tweets are sentiment analysed into 3 categories: Negative, Neutral & Positive.
- The probability of each is multiplied with the Tweet's retweet count to get a weighted score.
- Weighted & Individual scores are aggregated for each day using MapReduce.
- The aggregated sentiment data is merged with stock market data date-wise for analysis.
- Around 700 aggregated data points are available - (2 year span).
- Numerical data are min-max normalized to bring it to same scale of $[0 - 1]$.

Understanding Processed Data (cont.)

Table: Description of the Aggregated Tweet schema

Attribute	Description
Category	The industry sector name
Date	UTC date that the Tweet belongs to
Count	Total number of Tweets for a given category and date
Individual Negative	The aggregated negative confidence of the scored Tweets
Individual Neutral	The aggregated neutral confidence of the scored Tweets
Individual Positive	The aggregated positive confidence of the scored Tweets
Weighted Negative	The aggregated negative confidence of the Tweets weighted on retweet count
Weighted Neutral	The aggregated neutral confidence of the Tweets weighted on retweet count
Weighted Positive	The aggregated positive confidence of the Tweets weighted on retweet count
Negative counts	The number of negative Tweets for a given category and date
Neutral counts	The number of neutral Tweets for a given category and date
Positive counts	The number of positive Tweets for a given category and date

Correlation Analysis

- Spearman's rank correlation coefficient is used:

$$\rho = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{1}$$

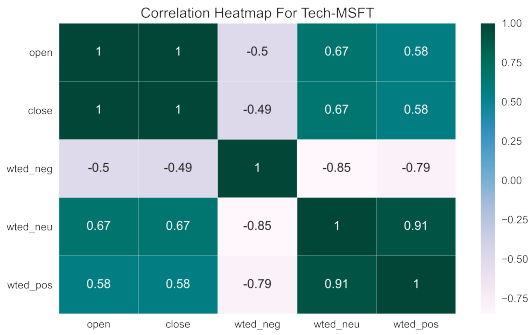
where:

  - ρ is Spearman's coefficient, $\rho \in \langle -1, 1 \rangle$
  - $cov(R(X), R(Y))$ is the covariance of the rank variables
  - $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the rank variables

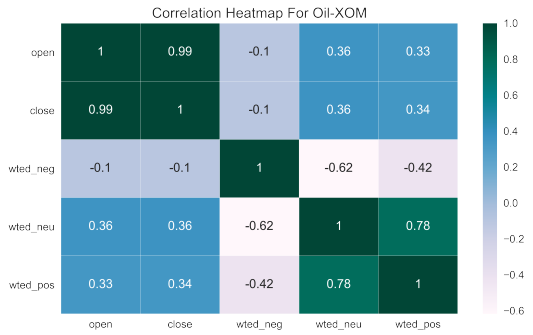
.

- Assesses monotonic relationships - irrespective of linearity (unlike Pearson's metric)
- Stock market and Tweet data may not be linearly related - thus Spearman's metric is better.

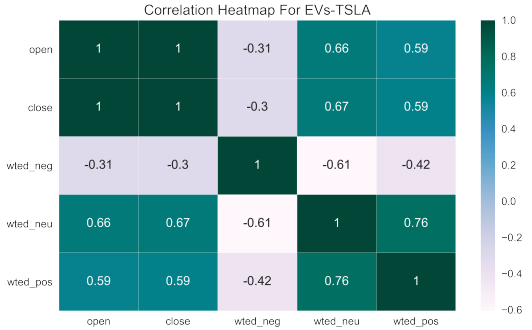
Correlation Analysis - Results



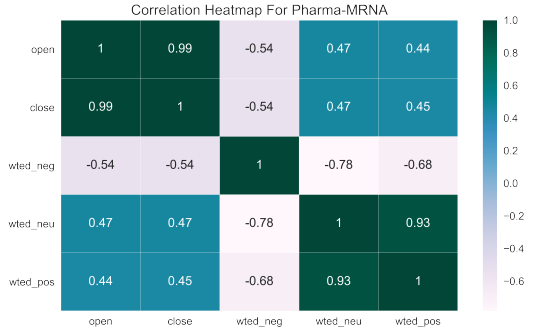
(a) MSFT



(b) XOM



(c) TSLA



(d) MRNA

Figure: Correlation heatmap between Tweets and Stock Market for different industrial sectors

Correlation Analysis - Discussion

- Tech Sector: Very good correlation observed. Tech market is volatile to mood expressed by social media.
- EV Sector: Very good positive correlation observed for Neutral & Positive sentiment. Moderate negative correlation for Negative sentiment - indicates market doesn't fluctuate as much if online mood is dull.
- Pharma Sector: Medium correlation observed for Neutral & Positive Tweet trends, good negative correlation for Negative sentiment - Pharma sector dips if online discussion is negative (comparatively).
- Oil Sector: Moderate correlation on Neutral and Positive, weak correlation on Negative. Market is not very susceptible to social media trends. Lower risk, comparatively stable.

Stock Market Prediction

- Since good correlation results are observed, market prediction can be done based on social media mood data.
- Next-day closing price prediction with social media sentiment scores (weighted).
- Boosted Decision Tree models used: XGBoost & CatBoost.
- Models are fine-tuned using Grid Search Algorithm.
- Separate regressor models are trained for each sector's dataset - with 85% data for training and 15% data for testing.

XGBoost Prediction - Results

(a) (NASDAQ: MSFT)

(b) (NYSE: XOM)

XGBoost Prediction - Results (cont.)

(c) (NASDAQ: TSLA)

(d) (NASDAQ: MRNA)

Figure: xgBoost Prediction Results

XGBoost Prediction - Metrics

Table: XGBoost Regressor Metrics

(a) (NASDAQ: MSFT)

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

(b) (NYSE: XOM)

RMSE	0.1914
MSE	0.0366
MAE	0.1420
R2 Score	-0.6285
Explained Variance Score	0.2131
Max Error	0.4236

(c) (NASDAQ: TSLA)

RMSE	0.0562
MSE	0.0031
MAE	0.0463
R2 Score	0.7086
Explained Variance Score	0.7555
Max Error	0.1630

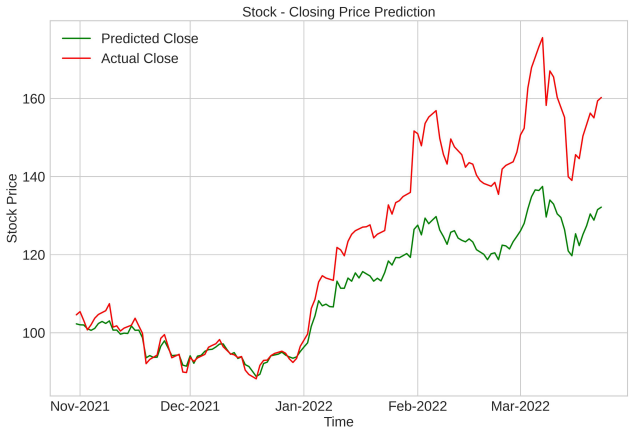
(d) (NASDAQ: MRNA)

RMSE	0.0366
MSE	0.0013
MAE	0.0255
R2 Score	0.8185
Explained Variance Score	0.8370
Max Error	0.1885

CatBoost Prediction - Results

(a) (NASDAQ: MSFT)

(b) (NYSE: XOM)



CatBoost Prediction - Results (cont.)

(c) (NASDAQ: TSLA)

(d) (NASDAQ: MRNA)

Figure: CatBoost Prediction Results

CatBoost Prediction - Metrics

Table: CatBoost Regressor Metrics

(a) (NASDAQ: MSFT)

RMSE	0.0480
MSE	0.0023
MAE	0.0383
R2 Score	0.8319
Explained Variance Score	0.8427
Max Error	0.1288

(b) (NYSE: XOM)

RMSE	0.1962
MSE	0.0385
MAE	0.1457
R2 Score	-0.7114
Explained Variance Score	0.1810
Max Error	0.4340

(c) (NASDAQ: TSLA)

RMSE	0.0711
MSE	0.0050
MAE	0.0562
R2 Score	0.5328
Explained Variance Score	0.6047
Max Error	0.2063

(d) (NASDAQ: MRNA)

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

Prediction - Discussion

- Both models perform remarkably well with good accuracy.
- Both models are able to capture the trend (rise/fall) accurately.
- Overall, the XGBoost model performs better in all cases except the Pharma dataset. It performs the best in Tech dataset with (83.81%) accuracy.
- CatBoost model however, is not very far behind. It performs the best in the Pharma dataset - (84.61%) accurate.
- Thus, both can be used for this task for good results, although XGBoost is slightly more reliable.

Conclusion

- A scalable, high-performance big-data pipeline was constructed to extract sentiment data from voluminous Twitter data and correlate it with stock market data.
- Conclusive proof was shown that stock market data has good correlation with social media (Twitter) sentiments.
- Furthered the line of research in this realm by taking into account past research's future work considerations.
- Provided a more fine-grained correlation analysis (sector-wise), rather than using index values like Dow Jones Industrial Average (DJIA) for analysis.
- Introduced two popular tree boosting models for this task, and highlighted their performance.

Future Work

- Consideration of other social media outlets viz. Facebook, Reddit etc. and also news articles - for sentiment analysis.
- More reliable stock market prediction with by including other features that affect market behaviour.
- Consideration of non-English Tweets (and written media), with similar NLP techniques to analyse them.
- Exploration of other market sectors for further generalization of our study.

References


Tianqi Chen and Carlos Guestrin. (2016). *XGBoost: A Scalable Tree Boosting System.* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, (785–794).


Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*


Gurcan, Fatih & Berigel, Muhammet, *Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges*, 2018 2^{nd} International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)


Kanavos, Andreas & Vonitsanos, Gerasimos & Mohasseb, Alaa & Mylonas, Phivos. (2020). *An Entropy-based Evaluation for Sentiment Analysis of Stock Market Prices using Twitter Data*


Chungho Lee, Incheon Paik, *Stock Market Analysis from Twitter and News Based on Streaming Big Data Infrastructure*


Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*


Loureiro D., Barbieri F., Neves L., Anke L. & Camacho-Collados J. (2022). *TimeLMs: Diachronic Language Models from Twitter.*


Mittal, A. (2011). *Stock Prediction Using Twitter Sentiment Analysis.*


Aparna Nayak, M. M. Manohara Pai and Radhika M. Pa, *Prediction Models for Indian Stock Market*


Dat Quoc Nguyen, Thanh Vu, Anh Tuan Nguyen, *BERTweet: A pre-trained language model for English Tweets*


Zhihao Peng, *Stocks Analysis and Prediction Using Big Data Analytics*, 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)

References (cont.)


Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. (2018). *CatBoost: unbiased boosting with categorical features.* In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, (6639–6649).


Pagolu, Sasank & Reddy, Kamal & Panda, Ganapati & Majhi, Babita. (2016). *Sentiment analysis of Twitter data for predicting stock market movements.*


Dharmaraja Selvamuthu, Vineet Kumar and Abhishek Mishra, *Indian stock market prediction using artificial neural networks on tick data*


Mehtab, Sidra & Sen, Jaydip. (2019). *A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing*. SSRN Electronic Journal. 10.2139/ssrn.3502624.


Vaanchitha Kalyanaraman, Sarah Kazi, Rohan Tondulkar, Sangeeta Oswal, *Sentiment Analysis on News Articles for Stocks*


Twitter-Roberta,`https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest`


Twitter-API, `https://developer.twitter.com/en/products/twitter-api`


Polygon-API, `https://api.polygon.io/v1/open-close/`


Apache Spark,`https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1`


Apache Kafka, `https://kafka.apache.org/downloads`


Yahoo-CMAK, `https://github.com/yahoo/CMAK`

Thank You

Thank You
Q & A