Correlation Analysis and Live Prediction of Stock Market using Social Media Trends

Nuzair Mohamed S	195001073
Sanyog Kave P	195001099

BE CSE, Semester 8

Dr. N. Sujaudeen
Supervisor

Project Review: 2 (25 March 2023)
Department of Computer Science and Engineering
SSN College of Engineering

1 Motivation

The value of a stock is influenced by various factors besides the traditional quantitative and qualitative analysis of stocks, such as financial statements. In recent years, social media trends have become a major factor affecting the value of a stock. This research aims to investigate the impact of these trends on stock value and to iteratively predict the price at which a stock is valued using live streamed data. The study focuses on Twitter and Reddit as the data sources and performs sentiment analysis on extracted text data to determine their correlation with stock values within the relevant industry sector and use batch and stream processing methods to develop and test an effective active-learning based prediction model. To efficiently handle the large amount of data involved, the study proposes using popular tools for real-time large-scale data analysis, such as Apache Spark and Apache Kafka.

2 Problem statement

Understanding and predicting the movement and the subsequent states of the stock market is difficult due to numerous factors, including economic climate, current trends, politics, and interest rates. Investors rely on multiple sources of information to make decisions about buying or selling stocks, including news events and social media trends. Social media trends, in particular, have become increasingly influential for amateur stock market investors as they tend to be regarded by them as highly credible sources of mass information. Immediate reactions to the stock market can be seen when incidents are reported or propagated using extreme sentiments on social media platforms. Hence it becomes essential to assess and quantify the effect of social media on the economics of different sectors and hence its influence on the stock market to effectively employ a prediction model. To analyze and infer the correlation, a large sample space with credible information is required. Since the considered data sources are of high volume with a high degree of velocity, deriving interesting patterns and inferences facilitate the need for a big-data architecture. Specific sectors like Tech, EVs, Finance and Health are used as the sample space for the analysis and prediction as there exists a growing innate correlation between these topics and the stock market. Hence predicting the stock values in these domains becomes the need of the hour with growing engagement and influence on the economic strength.

3 Architectural Design

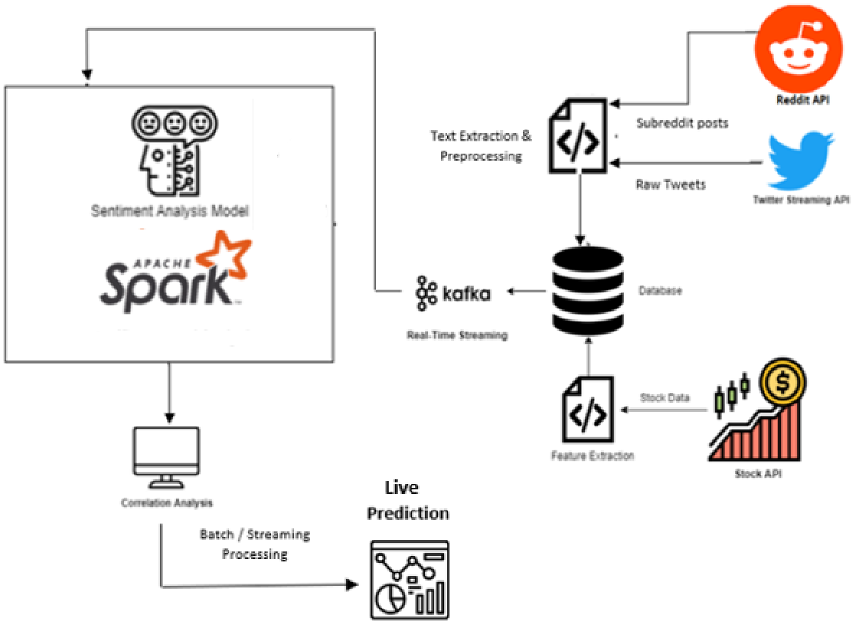


Figure 1: Proposed Architecture

We consider twitter tweets and reddit posts as source for the proposed system and analyse the relation between their sentiments and its influence on the corresponding stock entities.

4 Proposed system

Data from several APIs are collected separately and then preprocessed to preserve cleanliness of the data. Data are now integrated to be of a consistent schema and loaded into a database. Stock data are gathered using Stock API and required features are extracted and also loaded into the database. Then streaming of data is carried out using Kafka. The input stream of data is now fed to a pretrained sentiment analysis model. Sentiment analysis is carried out and the resulting sentiments are aggregated per category per day. Stock data is also aggregated in the similar manner by calculating the percentage of change of the value. Now these data are correlated to derive inferences about the correlation and all these processing of data is done using Apache Spark. The correlation results can be now used to decide whether prediction of stock using the sentiments got can be performed. The data can be batch or stream processed based on the requirement and effectively used to predict the price of a stock at various needed points in time (Live prediction).

4.1 Twitter Data Collection

The Twitter Streaming API (GET https://api.twitter.com/2/tweets/search/all) was used to collect the data from Twitter. Textual data was extracted from the raw data by removing emojis and the junk characters. The general developer API had limitations like dataset, time period and filter conditions due to which, the Twitter Academic Research access API had to be used. This API was used to query the data by providing keywords for each category of analysis over the span of August 2022 to January 2023. A Python script was written to collect the data, and generate a dataset of tweets labelled according to their category.

4.2 Reddit Data Collection

The Reddit API (GET https://api.pushshift.io/reddit/search/submission/) was used to collect data from the social media platform Reddit. A python script was used to extract data and do small cleaning from the subredditsusing the specific keywords chosen and saved in a CSV file according to their extracted category.

4.3 Stock Data Collection

The Polygon API (GET https://api.polygon.io/v1/open-close) was used to collect stock ticker data. The API accepted the type of ticker data required, and the date to get the data from. A Python script was used to collect Stock data for the span of August 2022 to January 2023, and was stored in a dataset categorized according to the sector of the companies queried. We get the opening, closing, high , low, premarket and aftermarket data of each stocks considered from this API.

4.4 Data Processing

The various APIs are used to collect tweets and reddit posts containing specific keywords or hashtags chosen based on their relevance to the stock market industries chosen.Sentiment analysis is performed on the tweets using the BERTweet NLP model[10], which outperforms previous state-of-the-art models. The output of the sentiment analysis is a 3-tuple of confidence values which is then converted to a single integer (-1: negative; 0: neutral; 1: positive) to represent the sentiment of the tweet. From the sentiment score for each data point that is calculated individually, a weighted average is taken. The category classifier and sentiment analyzer are set up on Spark Lambda Architecture, with the new live data fed to Spark using Kafka to simulate real-time analysis.

1. **Stock Approximation**
   The stock data collected form the polygon api had a lot of missing data on few days. The opening and closing value of the stocks for these missing days has to be approximately computed.We consider the opening, closing ,high and low of the day in which data is available before and the same for the day after and compute the average value for the missing days.

2. **Data aggregation**
   The combined twitter and reddit data are aggregated based on Category and Date along with the count of each sentiment for correlation analysis. Raw tweet data is thus processed using map-reduce methods to thus produce an overall sentiment of a market category on a given day. The stock values are also aggregated based on category and date to also be used for correlation analysis study. The percentage change of a stock for a given category on a given date is calculated using the opening and closing value as (close-open)/open and this acts as a representative of the statistics of the market of that category on that day. This aggregation task can be done as a real-time streaming task using Apache Kafka and Spark and this data is now used for further events.

4.5 Models used

1. **The BERT NLP Model:**
   BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language model developed by Google that is used for various NLP tasks such as sentiment analysis, text classification, and named-entity recognition. It has two versions: BERT base and BERT large, with the latter having double the number of layers compared to the base model. The input to the BERT encoder is a sequence of tokens, where each token is embedded with its position in the sequence and a segment identifier to distinguish between sentences. The final input to the BERT encoder is the sum of these three embeddings. The BERT model processes the input sequence by passing it through a series of self-attention layers and feed-forward neural networks in each block. The output of the BERT model can be used as features for various NLP tasks.

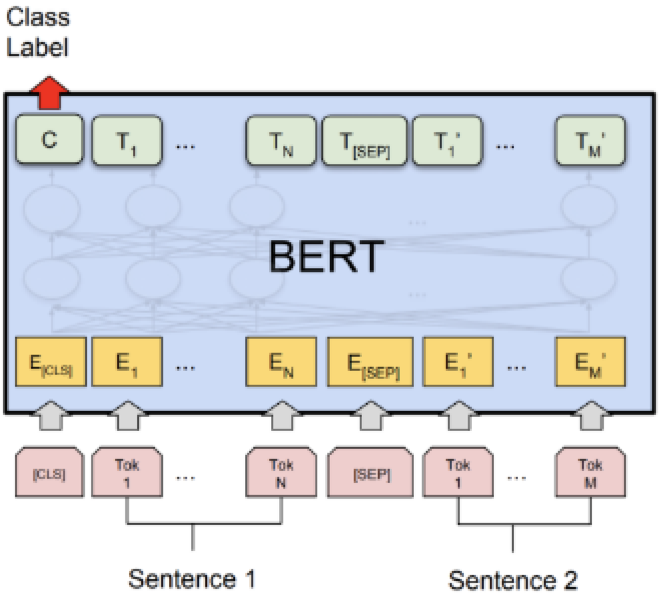
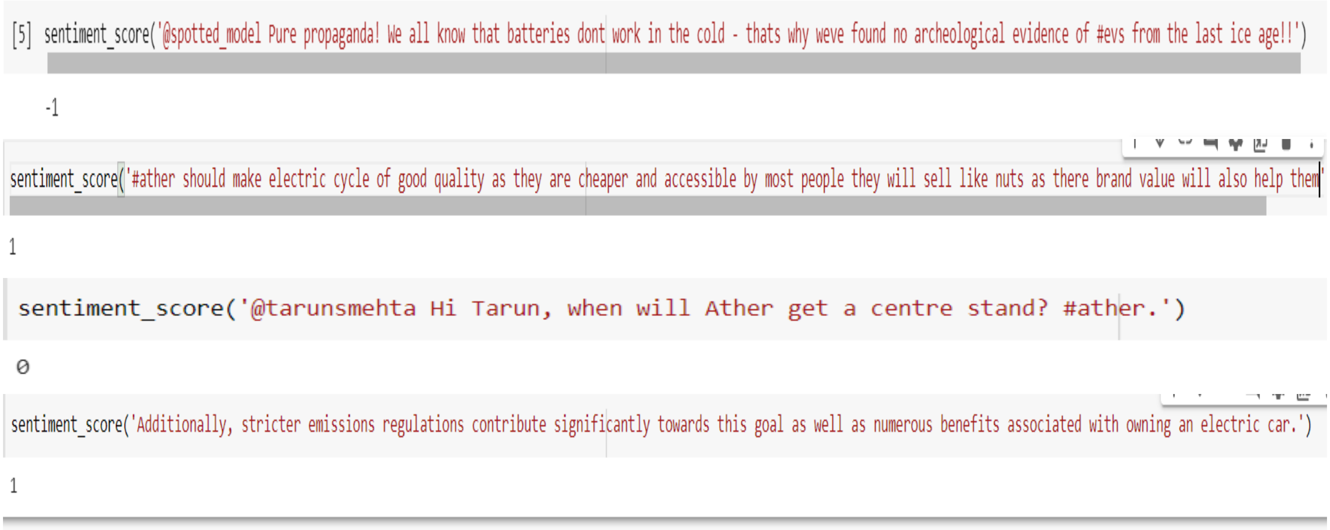


Figure 2: The BERT Sequence Classifier

2. **Sentiment Analysis:**
   Various models were developed to analyse the sentiment in a sentence.According to Nguyen et al.[10], the BERTweet model presented by them performs much better than existing models. This model (cardiffnlp/twitter-roberta-base-sentiment-latest) is being used to get the sentiment score for the collected data. The output is a 3-tuple values which is then converted to a single integer (-1: negative; 0: neutral; 1: positive) to represent the sentiment of the tweet. Few sample results

we got are as follows:



```
[5] sentiment_score('@spotted_model Pure propaganda! We all know that batteries dont work in the cold - thats why weve found no archeological evidence of #evs from the last ice age!!')

-1

sentiment_score('#ather should make electric cycle of good quality as they are cheaper and accessible by most people they will sell like nuts as there brand value will also help them'

1

sentiment_score('@tarunsmehta Hi Tarun, when will Ather get a centre stand? #ather.')

0

sentiment_score('Additionally, stricter emissions regulations contribute significantly towards this goal as well as numerous benefits associated with owning an electric car.')

1
```

Figure 3: Sentiment analysis done on sample tweet data

4.6 Statistical Correlation Analysis

The primary goal of this research is to identify if any correlation exists between data from social media sources like Twitter & Reddit and the stock market data in real-time. If there exists such a correlation, then we can quantify and visualize it. Correlation analysis is done with respect to sectors as the amount of influence that social media trends has on each sector might be contextually different. The Pearson correlation coefficient is a valid measure of linear correlation between two data sets. It is defined as the ratio between the covariance of two variables and the product of their standard deviations. This makes it a normalized measure that makes the result fall in the range of -1 to 1.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Figure 4: Pearson Correlation Coefficient formula

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

$$\rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2 - 1)}$$

Figure 5: Spearman Rank Correlation Coefficient formula

5 Techniques Used

5.1 Storage of collected data

The collected data is of large volume and hence needs a data streaming framework like Apache Kafka to efficiently process it. Hence the collected data has to be given to Kafka for which it is stored in an SQLite database. Data from reddit and twitter are stored together category-wise since they are extracted in the same format. Storing of data in SQLite database is done using a python script.

5.2 Data Ingestion using Apache Kafka

The collected data that now exists in SQLite database is fed to Kafka through producer script. The data are produced in micro-batches of size 500 into topics that are then consumed by the consumer script. This periodic streaming simulates a lives setup and facilitates uniform processing time to evaluate the performance.

5.3 Data Processing using Apache Spark

The incoming stream of data into topics is processed using a Spark application that facilitates real-time streaming and experimenting with correlation analysis. The PySpark interface is used to implement the Spark application. Usage of PySpark makes the integration of the application with BERT sentiment analysis model an easy task by using the PyTorch library.

5.4 Correlation Analysis

The sentiment towards a sector on social media and the tracked changes in stock market are observed to perform a correlation study. The aggregated data are merged on the basis of the Category and Date and normalized to perform scaling to observable ranges. Statistical correlation techniques like Pearson correlation coeeficient and Spearman correlation coefficient are applied and results are analyzed. This data can also be used to visualize and present the inferences using graph plotting libraries like Matplotlib.

6 Results

The social media data and stock data was collected as mentioned above and stored in the sqlite database as seperate tables.
Streaming of voluminous data using Kafka with the help of producer and consumer paradigm was set up.
Data are processed and aggregated using Spark application. The data streamed from Kafka is ingested and processed. Sentiment analysis and stock aggregation on Category and Date were carried out.
Correlation analysis is carried out based on the aggregated data for each category separately.

7 Impact to Society, Safety, Legal aspects

The whole system is based on the society as it analyzes societal emotions and relies on its herd mentality. The volatility and unpredictable nature of the stock market is always an intriguing topic that is noticed at large as it involves monetary aspects. Monetary safety can be ensured if the model provides very accurate predictions in real-time based on any social media outburst. There cannot be any legal complications as the proposed system is only a computational model and is always prone to errors.

8 Further Work

1. **Performing the Analysis:** The correlation analysis performed by Spark must be conducted and presented for all categories using appropriate visualization tools, in order to draw conclusions.

2. **Extrapolation of Results:** The results obtained thus far have to be extrapolated to identify and analyze if there are any other factors to weigh in that can further impact the prediction task.

3. **Live Prediction:** Create an active learning model from the data that can be used to iteratively predict the stock value trend in real-time and analyze the performance metrics

References

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

[2] Shashanka Venkatesh, Venkataraman Nagarajan, Vishakan Subramanian ,*Social media based Stock Market analysis using big-data infrastructure*

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

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

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

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean,*Efficient Estimation of Word Representations in Vector Space*

[7] Marsland, Stephen,*Machine Learning: An Algorithmic Perspective*, 2014

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding.*

[9] 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.*

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

[11] Fatih Gürcan, Muhammet Berigel, *Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges, 2018*