

Correlation Analysis - Stock Market & Social Media Trends

Nuzair Mohamed S 195001073

Sanyog Kave P 195001099

BE CSE, Semester 8

Dr. N. Sujaudeen

Supervisor

Project Review: 1 (10 February 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. The study focuses on News Articles, 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. 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 news and

social media on the economics of different sectors and hence its influence on the stock market. To analyze and infer the correlation, a large sample space with credible information is required. Since the considered data sources are of high volume and of high degree of velocity, deriving interesting patterns and inferences facilitate the need for a big-data architecture. Specific sectors like Telecom, Tech, EVs, Finance and Health are used as the sample space for the analysis as there exists a growing innate correlation between these topics and news articles.

3 Architectural Design

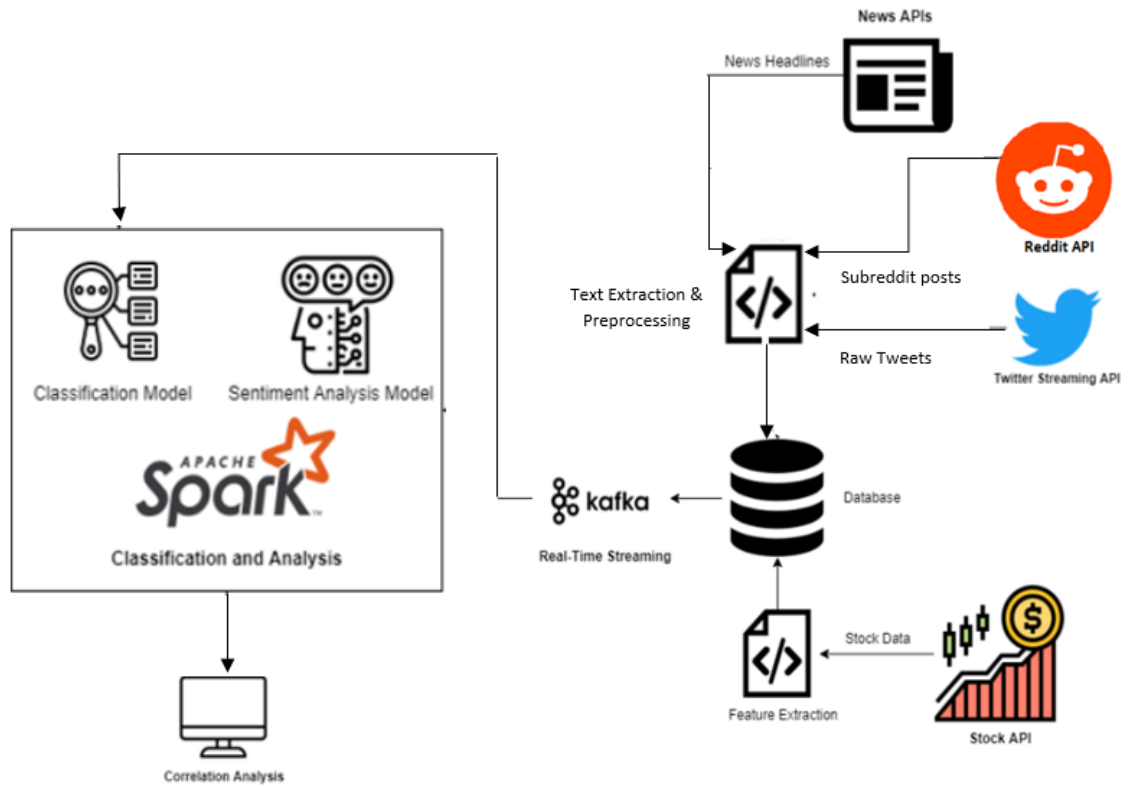


Figure 1: Proposed Architecture

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

3.1 News Data Collection

Three Python scripts were used to collect data from each of the APIs shown below and collect the combined data in a CSV file. The Headlines were labelled based on the the category they were extracted as.

1. NewsData: GET <https://newsdata.io/api/1>
2. NewsCatcher: GET <https://api.newscatcherapi.com/v2/search>
3. NewsApi: GET <https://newsapi.org/v2/everything?>

3.2 Twitter Data Collection

The Twitter Streaming API (GET <https://api.twitter.com/2/tweets/search/recent>) along with keywords for each category of analysis was used to collect tweets over the span of January and February of 2023. A Python script was written to collect the data, and generate a dataset of tweets labelled according to their category.

3.3 Reddit Data Collection

The Reddit API (GET https://www.reddit.com/api/v1/access_token) was used to collect data from the social media platform Reddit. A python script was used to extract data from the subreddits and saved in a CSV file according to their extracted category.

3.4 Data preprocessing

The various apis are used to collect tweets, news data and reddit posts containing specific keywords or hashtags chosen based on their relevance to the stock market industries chosen above. The collected data are then filtered using a machine learning model trained on a sample news headlines dataset to remove false positives and classify the rest upcoming data into relevant categories.

Sentiment analysis is performed on the tweets using the BERTweet NLP model, 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. 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.

3.5 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 January and February of 2023, and was stored in a dataset categorized according to the sector of the companies queried.

3.6 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.

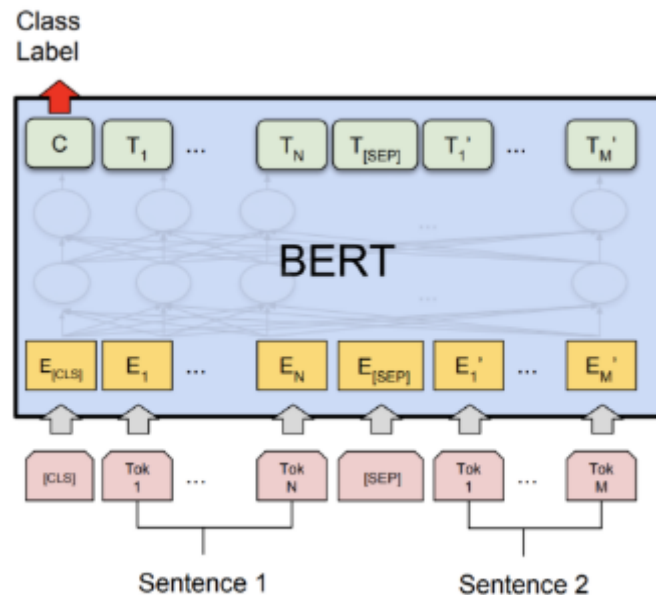


Figure 2: The BERT Sequence Classifier

2. The Classification Model:

The classification model, which is used to strip false-positive source data, is built using the BERT Sequence Classifier.

The model obtained a 84% accuracy on the test news data of approximately 2100 samples with the following Confusion Matrix (The range 0 - 4 represents the encoded values for the 5 different sectors):

	precision	recall	f1-score	support
0	0.91	0.78	0.84	50
1	0.91	0.88	0.90	122
2	0.82	0.87	0.85	63
3	0.96	0.82	0.89	131
4	0.56	0.82	0.66	55
accuracy			0.84	421
macro avg	0.83	0.83	0.83	421
weighted avg	0.87	0.84	0.85	421

Figure 3: Classifier confusion matrix

{'Telecom': 0, 'Health': 1, 'EVs': 2, 'Tech': 3, 'Finance': 4}

Text(33.0, 0.5, 'Ground Truth')

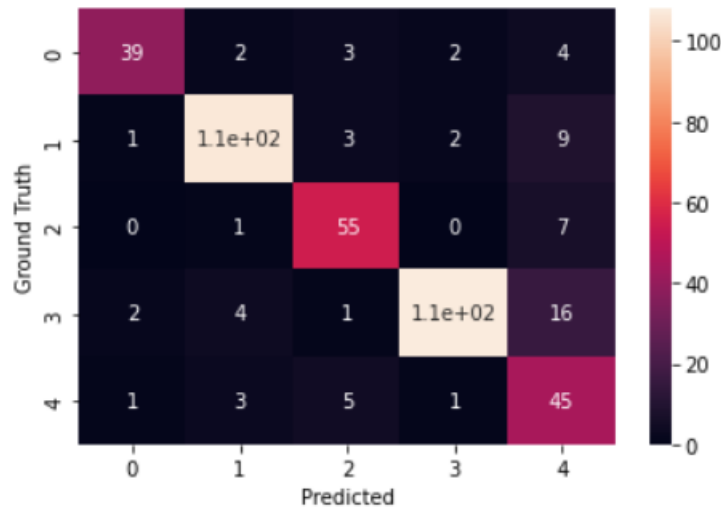
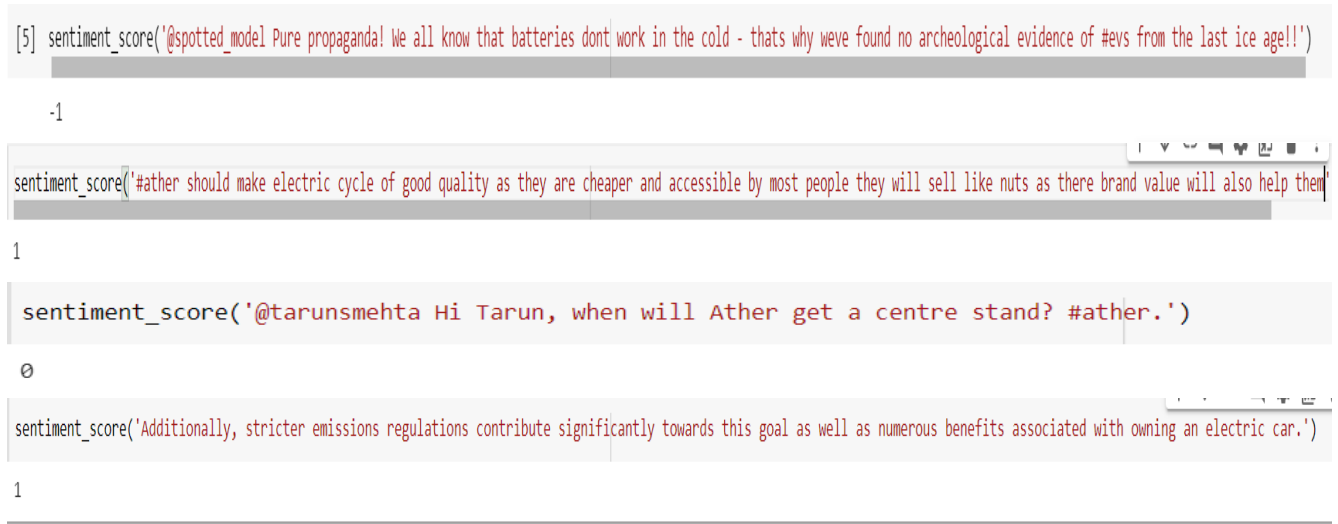


Figure 4: Fine tuning the BERT model using News Data

3. Sentimental Analysis:

Various models were developed to analyse the sentiment in a sentence. According

to Nyugen et al.[10], the BERTweet model presented by them performs much better than existing models. This model 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('@tarunsmehhta 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 5: Sentiment analysis done on sample tweet data

4 Literature survey

1) Lee et al. : It 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.

2) Shashanka et al. : Increase in volume of latest data from twitter and improved ML model.

Scope of improvement :

- News articles and platforms like Reddit also affect stock market.
- New huge sets of data from different sources as well as updated data-set from twitter.
- Integration of data from multiple sources require better ML model to improve classification and sentimental analysis.

5 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 live streaming of data is carried out using Kafka. The input live stream of data is now fed to a pretrained classification model that preserves only the data relevant to the sectors that are considered for correlation analysis. Sentiment analysis is carried out and the resulting sentiments are correlated with the trend in stock data to derive inferences about the correlation by Apache Spark.

6 Expected Outcomes

The main purpose of the entire analysis, is to discover a correlation between information from social media and news articles and stock market changes, if any. The expected outcome, therefore, is an accurate correlation analysis between social media trends, and the stock market variations in the related sectors and to infer if social media influences stock market.

7 Further Work

1. **Building a Kafka pipeline:** The data from SQLite database must be streamed using Apache Kafka to be then fed into Apache Spark.
2. **Configuring Apache Spark:** The Spark Lambda system needs to be configured to run the models, and to perform correlation analysis.
3. **Performing the Analysis:** The correlation analysis performed by Spark must be conducted and presented using appropriate visualization tools, in order to draw conclusions.

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*