Correlation Analysis and Prediction of Stock Market using Social Media Trends

Group ID: G5_8

Nuzair Mohamed S
Sanyog Kave P

Dr. N. Sujaudeen
Associate Professor

SSN College of Engineering, Chennai

March 25, 2023

Outline

- **Motivation**
- **Problem Statement**
- **Architectural Design**
- **Proposed System**
- **Results**
- **Expected Outcomes**
- **References**

Motivation

- In recent years, it is seen that stock values are highly influenced and impacted by social media events.
- Social media trends are increasingly influencing the stock valuations as seen from time to time.
- Due to this, we see frequent and unexpected stock market fluctuations more recently.
- Immediate reactions to the stock market can be seen when incidents are reported or propagated using extreme sentiments.

Problem Statement

- In a nutshell, we wish to analyse the correlation between social media trends & the stock market relating to specific sectors and its effect on their current stock valuations.
- To develop a big-data architecture that is capable of handling real-time streaming information from renowned social media platforms like Twitter, Reddit etc.
- 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.

Architectural Design

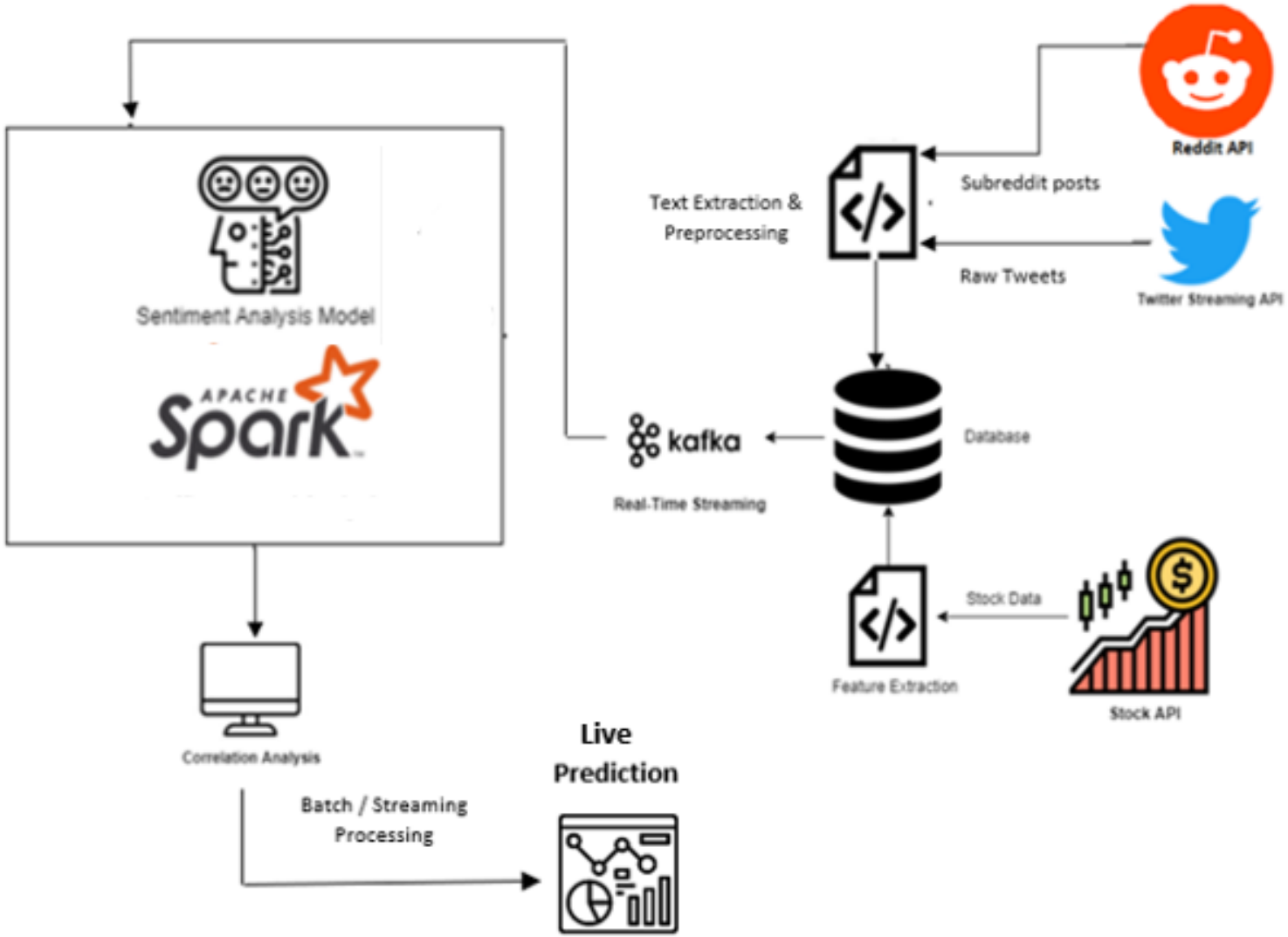

Figure: Proposed architecture of the system

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.
- 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 such as close, high, low value.

Comments received in 2nd review

- Complete Prediction task
- Performance analysis of Prediction modules
- Simulation of real-time prediction

Course of action

- Employed CatBoost and LightGBM prediction models
- Performance of each model for the attributes close, high, low for each stock is tabulated and compared.
- Summarize the entire project in a single python script that can simulate a real time prediction of close, high and low values of the stock for the next day by collecting the data for previous day.

Expected Outcomes

The main purpose of the entire analysis, is to unearth a correlation between information from social media through twitter and reddit and stock market changes, if any. The expected outcome, therefore, is an accurate correlation analysis betweeen social media trends, and the stock market variations in the related sectors and to infer if social media influences stock market. The system also aims at building an active learning model that can iteratively predict the stock valuations such as close, low and high.

References

- Chungho Lee, Incheon Paik, *Stock Market Analysis from Twitter and News Based on Streaming Big Data Infrastructure*
- Shashanka Venkatesh, Venkataraman Nagarajan, Vishakan Subramanian ,*Social media based Stock Market analysis using big-data infrastructure*
- Vaanchitha Kalyanaraman, Sarah Kazi, Rohan Tondulkar, Sangeeta Oswal, *Sentiment Analysis on News Articles for Stocks*
- Zhihao Peng, *Stocks Analysis and Prediction Using Big Data Analytics*, 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS).
- Kanavos, Andreas & Vonitsanos, Gerasimos & Mohasseb, Alaa & Mylonas, Phivos (2020), *An Entropy-based Evaluation for Sentiment Analysis of Stock Market Prices using Twitter Data*
- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean,*Efficient Estimation of Word Representations in Vector Space*

References - Contd.

- Marsland, Stephen,*Machine Learning: An Algorithmic Perspective*, 2014
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding.*
- 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.*
- Dat Quoc Nguyen, Thanh Vu, Anh Tuan Nguyen, *BERTweet: A pre-trained language model for English Tweets*
- Fatih Gürcan, Muhammet Berigel, *Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges, 2018*

THANK YOU