Social Media based Stock Market Analysis using Big-Data Infrastructure

Shashanka Venkatesh	18 5001 145
Venkataraman Nagarajan	18 5001 192
Vishakan Subramanian	18 5001 196

BE CSE, Semester 7

Dr. N. Sujaudeen
Supervisor

Project Review: 0 (26 October 2021)
Department of Computer Science and Engineering
SSN College of Engineering

1 Motivation

Stock market analysis has been a topic of great interest ever since public stock exchanges came into existence. It is an avenue that offers great profit, which by itself is a stimulus for most researchers who work in this realm. The task is also considerably demanding due to these returns as well as high randomness and various external influences affecting the current valuation of a stock.[7]

There are several factors that influence the value of a stock apart from the typical quantitative and qualitative parameters seen in fundamental analysis of stocks like balance sheets, income statements, cash flow statements etc. One of the most influential factors in recent years are news articles and social media trends.

The sentiments expressed by such social media posts & news articles cause a lot of volatility in the share market, leading to unexpected fluctuations of the stock value. A recent example is the inflation of Game Stop's stock, which was primarily triggered by an online Reddit community to go against the predictions made by Wall Street experts.

The aim of this research is to analyse and understand the effect of such articles over the value of a stock at any given time. To analyse such voluminous amounts of data in an efficient & effective manner, we propose to build an architecture that makes use of popular tools for analysing large-scale data at real time, like Apache Spark & Apache Kafka.

Keywords: *Stock Market Analysis, Big-Data, Social Media, Sentiment Analysis, Machine Learning, Twitter, Apache Spark, Apache Kafka.*

2 Problem Statement

In a general sense, stock market analysis aims to determine the future movement of the stock value of a financial exchange. An accurate prediction of such movement leads to more profits for an investor. Predicting how the stock market moves is a challenging issue due to many factors involved in it, like interest rates, economic growth, current trends, politics, etc.

Stock market experts use various sources of information to determine whether or not to buy/sell stocks of any company. Two such sources that are widely used (especially in recent years) are news events & social media trends.

Our goal here is to test this very hypothesis. We wish to understand and analyse the statistical correlation between these news trends and its effect on the relevant industry's stock price valuations.

To correctly evaluate the correlation, we require a large sample space with properly collected data points from various trustworthy sources of information. Stock data, current news & social media trends tend to be voluminous with a high degree of velocity. To appropriately wrangle such copious amounts of raw data to extract valuable information out of it requires the usage of a big-data architecture. The features being explored here also perfectly conform to the 5V's that generally define big data - volume, variety, veracity, velocity and value.

Processing of big-data can be categorized into three distinct types: batch, real-time and hybrid processing. Batch processing is a well-organized technique of processing high volumes of data, and is not time limited. Real-time processing involves continuous input, processing and reporting of data, and highly focuses on achieving the lowest latency. The hybrid paradigm employs batch as well as real-time processing, wherein the results from both are analyzed & queried together and then combined & evaluated.

3 Literature Survey

3.1 A Practical Perspective

The stock market hasn't been left out in the wave of data science and machine learning that has swept the entire world. It has become a prime focus of various ML and AI researchers given the volatile and unpredictable nature of the stock market. Stock market prices are basically influenced by supply and demand from the market. The rate at which stocks rise or fall in value depends on how the stock's demand is throughout. The stock market, being partially regulated, completely lies in the shareholder's hands to properly predict and buy/sell the shares.

Some factors that affect stock prices are:

- Random Speculation
- Government Policies
- News
    - Price Hike
    - Fraud detection
- Trends
- Seasonal(s)
- Similar industry growth/decline

Our main idea is to analyse the correlation between stock valuation and one of the factors that exerts a significant degree of influence on it, especially in recent years.

3.2 Previously Explored Ideas

Lee et. al.'s[1] 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.

However, the paper falls short on one main point - lack of statistically significant data. The paper details that they used a sample of 1000 news articles to create a classifier, which was appropriate. They were able to obtain a 77% accuracy on the classification of Tweets into different sectors. The source of the statistical insignificance comes from the fact that they only used 100 Tweets overall, with 20 Tweets per category in order to test out their hypothesis.

This seems to be a statistically insignificant volume of data, which reduces the confidence in their conclusion that the correlation is insignificant.

Another significant reason for exploring this area is the fact that the internet is a rapidly growing community. Even though a good portion of the world's population was online in 2017, there has been an increase of nearly 70 million users on Twitter between 2017 and 2021. Recently, the cryptocurrency sector has seen a major impact from Tweets regarding it, mainly from Elon Musk, which have immediately triggered varying shifts in public opinion, leading to sudden change in cryptocurrency stock valuations. This gives yet another reason to explore the possibility that there could be statistically significant correlation between Tweets and stock prices.

Selvamuthu et. al[6] also express the opinion that sentiment analysis of the opinions of renowned personalities might help get an extra edge in stock price prediction, since their opinions are known to affect stock prices.

Nayak et. al[5] have done significant work on sentiment analysis of Twitter data for stock price prediction using the NLTK package and trend analysis of stock data obtained from Yahoo Finance for three sectors: Banking, Mining & Oil. They have also put forth algorithms to calculate closed price trends for each day, to check continuous days up/down for a stock, and an algorithm to combine sentiment with historical data. A correlation model also has been described in their research work.

Kalyanaraman et. al's[2] research tries to perform such an analysis purely using news articles. They considered 11 companies under the National Stock Exchange (NSE), and considered 100 news articles for each company. They manually pre-processed the articles to remove irrelevant ones. This further reduced their dataset to 50-60 articles per company. They then manually labelled the sentiment of the articles as being positive or negative. They used a custom sentiment dictionary and vectorized the news articles to build a classifier using Linear Regression and Gradient Descent. Their model predicted the sentiment of articles with around 60% accuracy and the direction of stock price movement (positive or negative) with an impressive 81.92% accuracy.

Kanavos et al.[7] have used the Stanford Core NLP library to perform text preprocessing along with Naive Bayes classifier to perform sentiment analysis on Twitter data for stock price predictions. Their classification was done on 6 dimensions, which were Alert, Calm, Happy, Kind, Sure & Vital.

3.3 Exploring Technologies Involved

The initial step in building a model that finds the required correlation, is to classify real-time Twitter stream data into its relevant sectors. This involves three main steps:

1. Mapping news articles to word-vectors.
2. Learning a mapping from word vectors and the corresponding sector of the news article.
3. Using the learned classifier to classify tweets.

Mikolov et al.'s[8] research explores the idea of mapping words to the vector space. The crux of their work being converting words of a sentence to numerical vectors in an n-dimensional space, so that it can be used for further classification. Previous work done in the area were Neural Network Language Model (NNLM) and Recurrent Neural Net Language Model (RNNLM). These earlier methods showed few examples of similar words in the tabular form while Word2Vec model prepared a comprehensive test set which has five semantic relations and nine syntactic relations between words. Essentially this allows us to obtain a vector mapping of the headline under consideration.

Going further, we need to map this vectorized headline to a sector. We need an ML model that can map vectors to an encoded set of sectors, and be able to classify any new vector to one of the sectors. Marsland[9] confirms that the task is perfect for a Single/Multi-layer Perceptron.

On the other hand, we need a scalable & robust architecture to handle big-data collected from the various information sources. Fatih et. al[4] explains the lifecycle associated with real-time big-data processing and highlights the use-cases of different big-data tools. From their research, we chose to go with Apache Kafka for our Data Ingestion module & Apache Spark/Spark Streaming to handle batch and stream processing of data. The paper also mentions the different challenges associated with unstructured data, which should be kept in mind while constructing our architecture.

Peng's research work[3] gives an overview of machine learning with Apache Spark for stock market analysis. The paper also proposes a generic pipeline for big-data architectures. It also explains on how to pre-process data using PySpark.

4 Proposed System

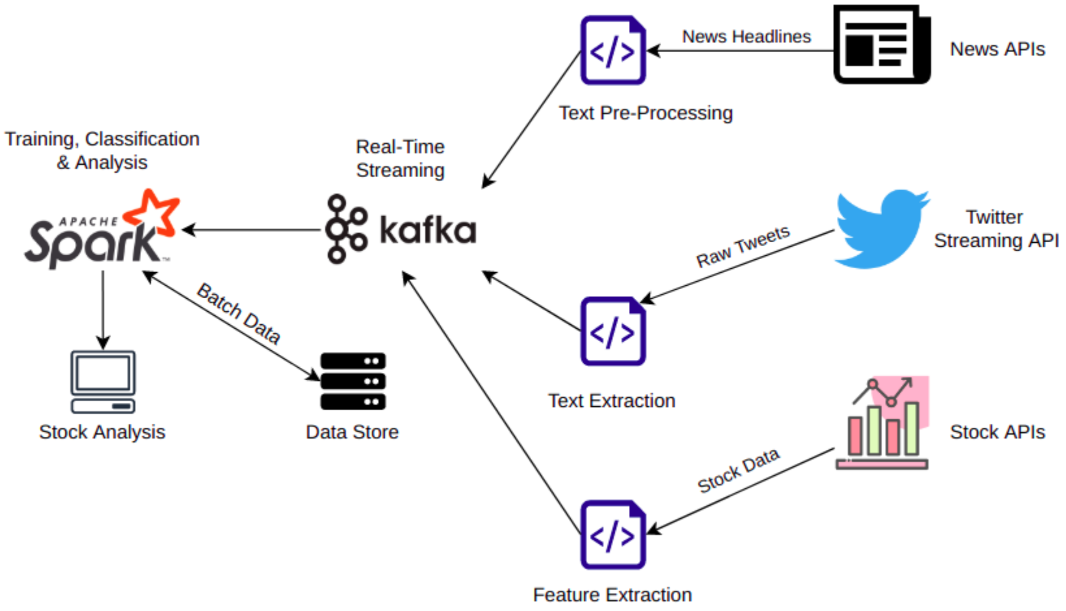

Figure 1: Proposed Architecture

4.1 Twitter Data Classification

News headlines data is used to train a machine learning model to classify text phrases onto their relevant sectors. For example, a tweet with the phrase: ”Invest more in BitCoin!” would be classified onto the cryptocurrency sector.

Real-time tweet data is then aggregated using the Twitter Streaming API through Apache Kafka and then classified. Following this, tweets under each subclass are independently processed using machine learning and text recognition to perform sentiment analysis to understand if the tweet is positive, negative or neutral. For this task, we propose to use powerful NLP tools provided by libraries like NLTK, SpaCy etc.

The outcome from the above sentiment analysis would then be correlated onto its sectors’ current stock market value. We can hence quantify the effect of social media on the particular sector’s economy.

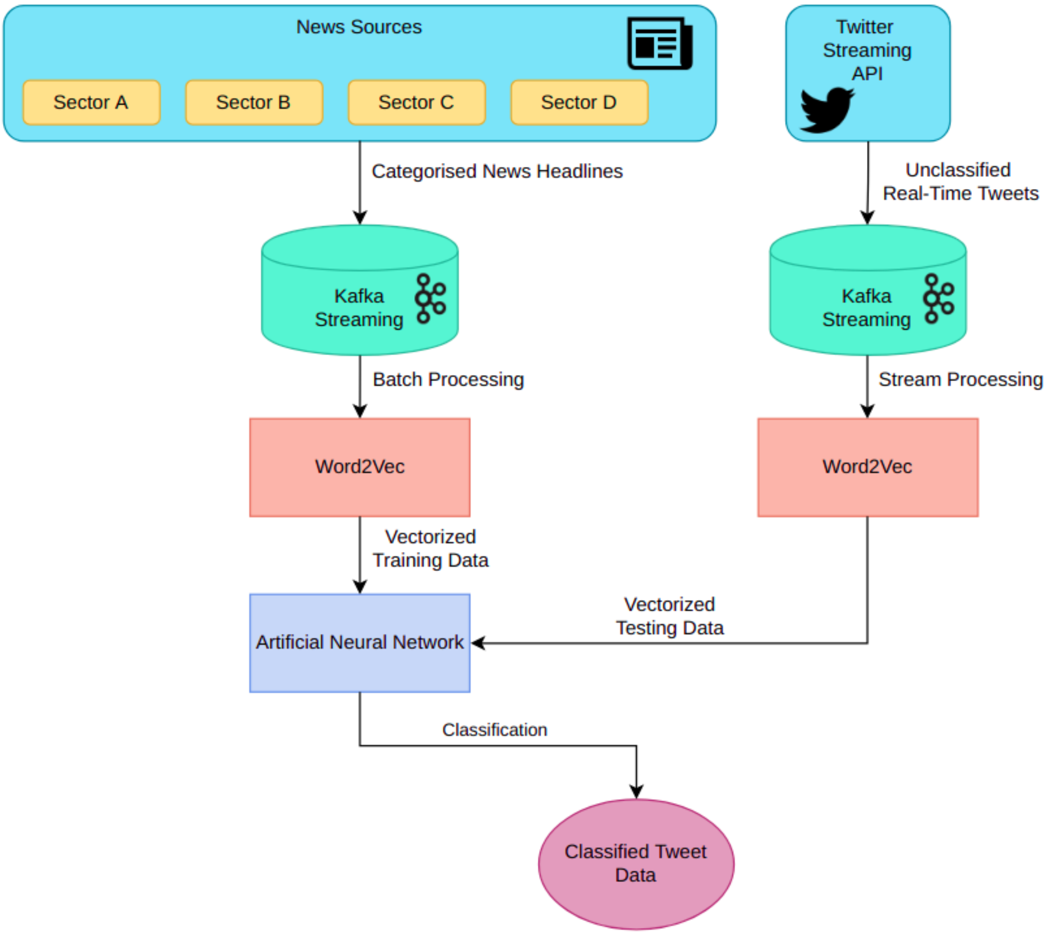

Figure 2: Pipeline For Sentiment Analysis

4.2 Stock Data Analysis

Raw stock data is collected from public stock APIs at frequent intervals at real-time using Apache Kafka, which is then pre-processed. Following this, the pre-processed data is trend analysed using Apache Spark to understand shareholders' current mood towards that sector's stock. This trend analysis is then compared with the trends & public moods seen from the Twitter sentiment analysis, for further statistical studies.

4.3 Statistical Correlation Analysis

The main objective of our research is to identify whether there exists any correlation between Twitter data & stock market data in real-time, and if so, quantify the correlation between them in different sectors.

Correlation analysis needs to be done sector-wise, since social media might have a varying degree of influence over each sector - a logical example would be the cryptocurrency market, which is highly volatile with respect to social media sentiments. This level of influence may not be seen with respect to other sectors like the entertainment or healthcare sector.

References

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

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

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

[4] 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)

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

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

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

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

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