WatchTrade
Company Stock Movement Modeling and Tracking

Harsh Dubey hd2225@nyu.edu
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

Samvid Zare sz3369@nyu.edu
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

Yulu Qin yq810@nyu.edu
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

Abstract

The project aims to develop a comprehensive and unified company profiler that integrates data from various sources to provide a more comprehensive view of a company's performance and outlook. The company profiler will help to extract insights and make predictions about a company's stock performance. The advanced technical indicators and performance predictions generated by the company profiler will be useful for domain experts, such as investors and financial analysts, who can use these insights to inform their decision-making. Additionally, the company profiler can be extended to provide custom performance dashboards that display the key insights and predictions in a user-friendly format, making it easy for domain experts to quickly understand a company's performance and outlook.

Keywords: Spark, Scala, HDFS, Data Infrastructure, Stocks

Datasets: Historical Stock Indicators, Public Conversations, Financial News

Platform: NYU Dataproc

1. Introduction

What would it take to build a company profiler like Yahoo Finance but much more Unified? The project idea is to work on a pipeline that uses heterogeneous data sources such as Stocks, Conversations and News to perform company profiling and stock movement analysis. This involves exploratory data analysis, feature engineering and extraction and modeling to build more advanced technical indicators which can be used by domain experts. The raw features derived from the data can also be fed into Machine Learning models to predict stock closing prices and performance indicators. All of the data modeling and profiling can be used to build a custom performance dashboards displaying insights about a particular asset.

Stock market profiling needs a lot of work and a team of data scientists, software engineers, and subject matter experts. Exploratory data analysis, feature engineering and extraction, and the creation of machine learning models are all tasks that belong to the data scientists. The pipeline and unique performance dashboards are constructed by the software engineers. The domain specialists advise on the most pertinent data to gather and examine, and they are in charge of deciphering the outcomes of the data modeling and profiling. In order for the team to work on improved strategies

and the application of actionable insights, this project can reduce the efforts of the entire team. Overall, although this project requires a significant investment of time, resources, and expertise to develop, it can provide a comprehensive and unified company profiler that provides valuable insights and predictions for domain experts.

2. Background And Motivation

2.1 Understanding the Users of the Project

The Project is a tool or service that is designed to help equity investors, day traders, and hedge funds with their trading activities. It is designed to be used by users of all levels of experience, from novices who are just starting out in the world of trading to seasoned investors who have a lot of market knowledge.

From locating prospective investment possibilities to carrying out deals and controlling risk, the Project seeks to help traders with a wide range of duties. For the purpose of assisting traders in making wise judgments and keeping track of market trends, it might provide features like real-time market data, analysis tools, and trading alerts. The Project's overall objective is to offer a thorough and user-friendly platform that may assist traders of all experience levels in enhancing their performance and achieving their financial objectives.

2.2 Understanding the Beneficiaries of the Project

The analytics provided by this project can be useful for all of the users mentioned above, including equity investors, day traders, and hedge funds, to help them manage their portfolios more effectively. Users can make more informed investment decisions by using the project's data and analytical tools, which can offer insights into market patterns and conditions. Users may be able to employ analytics, for instance, to monitor the performance of certain stocks, industries, or the entire market and to spot prospective buying or selling opportunities.

The insights produced by this initiative can be helpful for businesses looking to learn more about the mood surrounding their stock in addition to individual investors. Companies may better understand how their stock is seen in the market and how it is likely to perform in the future by examining data on how their stock is traded and the opinions of investors. Companies may find this information useful as they plan their operations and financial investments. Overall, the analytics from this project can be a valuable resource for a wide range of users, providing valuable insights into market conditions and investor sentiment.

2.3 Importance of the Project

According to Tesar and Werner (1995), the average U.S. portfolio remains strongly biased toward domestic equities. The project aims to provide users with a comprehensive view of the various channels that can affect the price of a company's stock. This includes information from social media and other internet sources, financial data and analysis, news about the business and its sector, and company- and industry-related news. The initiative intends to give users a unified perspective of the different elements that can affect the price movement of a company's shares by combining all of this information onto a single platform.

Users might be able to observe, for instance, how a company's stock price is affected by its financial performance and news releases, as well as how investor behavior is influenced by social media sentiment. Users may benefit from this by better comprehending the larger context in which a stock is trading and by being better equipped to make investment decisions. The project may also include tools and analyses to assist users in seeing trends and patterns in the data, giving them knowledge of how various channels interact and affect the price of a company's stock. Overall, the

goal of the project is to provide users with a comprehensive and easy-to-use platform that can help them stay informed and make better investment decisions.

3. Understanding Data and its Sources

3.1 Historical Stock Indicators

3.1.1 Understanding Data

This dataset contains historical pricing data for a collection of stocks. Each stock's daily closing values are included in the data, along with a number of indicators that show how these assets' prices have changed over time. Indicators are calculated numerical traits that are used by traders and investors to better understand a stock's behavior. They are based on pricing data. Marjanovic (2017)

The moving average, which displays the average price of a stock over a specified period of time, and the relative strength index, which determines if a stock is overbought or oversold, are two examples of indicators that might be derived using the attributes in this dataset. These indicators can offer useful information about how a stock's price changes, enabling consumers to choose whether to purchase or sell with greater knowledge. The dataset may also be used to discover additional features like trading volume, which shows how much of a stock has been traded over a specific time period. The dataset's overall goal is to arm users with a plethora of knowledge regarding stock performance across history, enabling them to better comprehend and evaluate these assets.

3.1.2 Data Sample

Date	Open	High	Low	Close	Volume
1962-01-02	0.6277	0.6362	0.6201	0.6201	2575579
1962-01-03	0.6201	0.6201	0.6122	0.6201	1764749
1962-01-04	0.6201	0.6201	0.6037	0.6122	2194010
1962-01-05	0.6122	0.6122	0.5798	0.5957	3255244
1962-01-08	0.5957	0.5957	0.5716	0.5957	3696430

Figure 1: Historical Stock Indicators Data Sample

3.2 Public Conversations

3.2.1 Understanding Data

This dataset includes discussions and comments about multiple stocks that have been posted in the public domain. This may include discussion boards, social media sites, blogs, and other websites where people discuss stocks and share their thoughts. The dataset contains the content of these conversations and remarks in addition to metadata, such as the date and time at which they were posted, the individual who made them, their tone, and any other pertinent details.

In order to enable users to see what people are saying about various firms and their stocks, this dataset aims to provide a record of public discussions and comments regarding stocks. Investors that are looking to understand market sentiment and get knowledge about how the general public views various stocks may find this to be helpful. Overall, the dataset is intended to provide a

comprehensive view of public discussions and comments about stocks, offering valuable insights for investors and traders.

3.2.2 Data Sample

	message_body	sentiment	timestamp
0	$FITB great buy at 26.00...ill wait	2	2018-07-01T00:00:09Z
1	@StockTwits $MSFT	1	2018-07-01T00:00:42Z
2	#STAAnalystAlert for $TDG : Jefferies Maintains with a rating of Hold setting target price at USD 350.00. Our own verdict is Buy http://www.stocktargetadvisor.com/toprating	2	2018-07-01T00:01:24Z
3	$AMD I heard thereâ€™s a guy who knows someone who thinks somebody knows something - on StockTwits.	1	2018-07-01T00:01:47Z
4	$AMD reveal yourself!	0	2018-07-01T00:02:13Z
5	$AAPL Why the drop? I warren Buffet taking out his position?	1	2018-07-01T00:03:10Z
6	$BA bears have 1 reason on 06-29 to pay more attention https://dividendbot.com?s=BA	-2	2018-07-01T00:04:09Z
7	$BAC ok good we're not dropping in price over the weekend, lol	1	2018-07-01T00:04:17Z
8	$AMAT - Daily Chart, we need to get back to above 50.	2	2018-07-01T00:08:01Z
9	$GME 3% drop per week after spike... if no news in 3 months, back to 12s... if BO, then bingo... what is the odds?	-2	2018-07-01T00:09:03Z
10	$SBUX STRONG BUY!	2	2018-07-01T00:09:26Z
11	$SNPS short ratio is 2.17 at 2018-06-15 and short % to float is 1.42% http://sunshineavenue.com/stock/SNPS/ via @sunshineave	-2	2018-07-01T00:09:36Z
12	$NFLX price squeezing,perfect place for an option straddle near the supporting trend	2	2018-07-01T00:12:58Z

Figure 2: Public Conversations Data Sample

3.3 Financial News

3.3.1 Understanding Data

This dataset includes various financial news published by authorized news providers. This includes information from the news and media regarding stocks and the businesses they stand for. The dataset contains a title and main news content along with the metadata, including the date and time of publication, the name of the publisher, and any other pertinent details.

In order to enable users to view what the media and other experts are saying about various firms and their stocks, this dataset aims to provide a record of the news regarding stocks that have been published in the news domain. Investors who are interested in learning about the most recent market happenings and how various stocks are being reported by the media may find this to be helpful. Overall, the dataset is intended to provide a comprehensive view of the news about stocks, offering valuable insights for investors and traders.

3.3.2 Data Sample

id	ticker	title	category	content	release_date	provider	url	article_id
221520	NIO	NIO leads consumer gainers Origin Agritech only loser	news	Gainers NIO NYSE NIO 14 Meritor NYSE MTOR 13 Eastman Kodak NYSE KODK 8 Village Farms International NASDAQ VFF 8 The Lovesac NASDAQ LOVE 8 Losers Origin Agritech NASDAQ SEED 7	31-12-2019	Seeking Alpha	https://seekingal	2050524
221521	NIO	Beyond Meat tops consumer gainers NIO and Eastman Kodak among losers	news	Gainers Beyond Meat NASDAQ BYND 7 Purple Innovation NASDAQ PRPL 6 Losers CTI Industries NASDAQ CTIB 18 Eastman Kodak NYSE KODK 10 Celsius Holdings NASDAQ CELH 7 e l f Beauty NYSE ELF 7 NIO NYSE NIO 6	07-01-2020	Seeking Alpha	https://seekingal	2054201
221522	NIO	U S Auto Parts Network leads consumer gainers NIO leads the losers	news	Gainers U S Auto Parts Network NASDAQ PRTS 8 Nova LifeStyle NASDAQ NVFY 5 Losers NIO NYSE NIO 8 Eastman Kodak NYSE KODK 6 Pyxus International NYSE PYX 6 The Lovesac NASDAQ LOVE 5 Planet Green NYSEMKT PLAG 5	02-01-2020	Seeking Alpha	https://seekingal	2051319

Figure 3: Financial News Data Sample

4. Hypotheses

To make sure that our infrastructure for stock market movement analysis works as expected, there are certain hypotheses that need to be proved correct to ensure the goodness of the system. So, let's look into these hypotheses one by one.

- **Close Price Prediction**:

4.0.1 Theory

Close price prediction suggests that there should be a way to predict the close price of the stock from other indicators or features. Close price is one of the indicators that are only available after the market has closed for a day. This means that given certain indicators/features of the day.

4.0.2 Importance of the Hypothesis

Close price is one of the most important indicators. Almost all the technical indicators that exist in the domain are in some form a transformation or aggregation of close price. This suggests that we cannot understand the market and make predictions without close prices. Given that it becomes available after the market closes, can we design a way in which we can predict the closing price and hence predict the technical indicators which can help us make informed decisions in advance?

4.0.3 Measure of Correctness

We can measure the correctness of the hypothesis if we are able to design a model that can use existing features/indicators to predict the closing price. Also, the distribution/trend of the closing price follows the real closing price closely.

- **Technical Indicators Prediction**:

4.0.4 Theory

Technical Indicators prediction suggests that there should be a way to find advanced technical indicators that are used in the market to understand stock movement using the information or features we have after feature engineering. For ex - Technical Indicators like RSI, Bollinger Bands, and PriceVolume should be possible to predict.

4.0.5 Importance of the Hypothesis

Technical Indicators are one of the most important aspects of stock market engineering and measurement. They are used to look into those aspects of the market which are not directly visible from raw market data. This gives great insights that are used by market experts to study the market, make predictions, perform transactions, and take action. Given that if we are able to engineer these indicators using the information we have, we will be able to measure the market, its movement, and its sentiment just by looking at these indicators.

4.0.6 Measure of Correctness

We can measure the correctness of the hypothesis if we are able to design a way to engineer these market technical indicators and prove that they follow the usual trend as theoretically suggested.

- **Correlation Between Public Conversation And Stock Movement**:

4.0.7 Theory

With the development of modern machine learning and deep learning models, public sentiment score generated by such systems can accurately reflect public attitudes on the stock market. Thus, with such public sentiments, we hypothesize that the sentiment of a public conversation over stocks per day generated from the conversation data-set has an impact on the oscillation of stock prices per day in the same time frame Pagolu et al. (2016). A positive correlation between sentiment and price is expected.

4.0.8 Importance of the Hypothesis

With the rise of social media, people have been posting vast amounts of information on a variety of topics. Considering the scale, speed, and timeliness of public discussions on social media over stocks, researchers have become increasingly interested in finding out what role financial-related conversations may play in financial models. Public sentiment is an important indicator of public attitudes and preferences over stocks. If a strong correlation can be observed, public conversations can be considered as an important predictor of stock price. Given the amount and accessibility of public conversations as well as the development of modern large language processing models, predicting stock price with public sentiment could be feasible and valuable.

4.0.9 Measure of Correctness

We can measure the correctness of the hypothesis if we are able to calculate the average sentiment of a sample stock over a certain time frame and map with its price variations. We can plot the oscillations and observe their correlations or extract out points and implement a correlation analysis.

- **Correlation Between News And Stock Movement**:

4.0.10 Theory

As we've considered user view utilizing the public conversation data-set, real-time information from news channels can provide valuable insight about stock price movement from the professionals' point of view. Seif et al. (2018) suggested market news and reports can reflect the pulse of the market and mirror the events occurring on the market during a stock market session. Considering the diverse contents provided by public news, including description and analysis of past political and economic events and their impacts, summary of the past stock performances, comments and prediction of future price movement made by specialists and financial critics, we hypothesize that public news conversation has a correlation with stock price movement Wang et al. (2018).

4.0.11 Importance of the Hypothesis

To further understand the relationship between public news and stock prices, we can use the same approach as we plan for the public conversation data-set– group the news by stock names and dates and generate an average sentiment score. However, since deriving features from news dataset cannot be directly achieved by sentiment analyzer due to large size of news content per stock and there is no support for deep learning model training which proposes challenges for sentiment analysis and related tasks, we chose number of news as the alternative indicator. Though merely number of mentions from the news without information about their sentiments may not always give a complete picture about professionals' attitudes, it can still reflect the popularity and public attention over a particular stock and ultimately yields insight on how the stock price may vary on a daily basis.

4.0.12 Measure of Correctness

We can measure the correctness of the hypothesis if we are able to calculate the number of news of a sample stock over a certain time frame and map with its price variations. We can plot the oscillations and observe their correlations or extract out points and implement a correlation analysis.

5. Architecture

Let's look into how we can design an architecture that we can use to carry out the tasks needed to design such infrastructure/pipeline.

5.1 Design Diagram

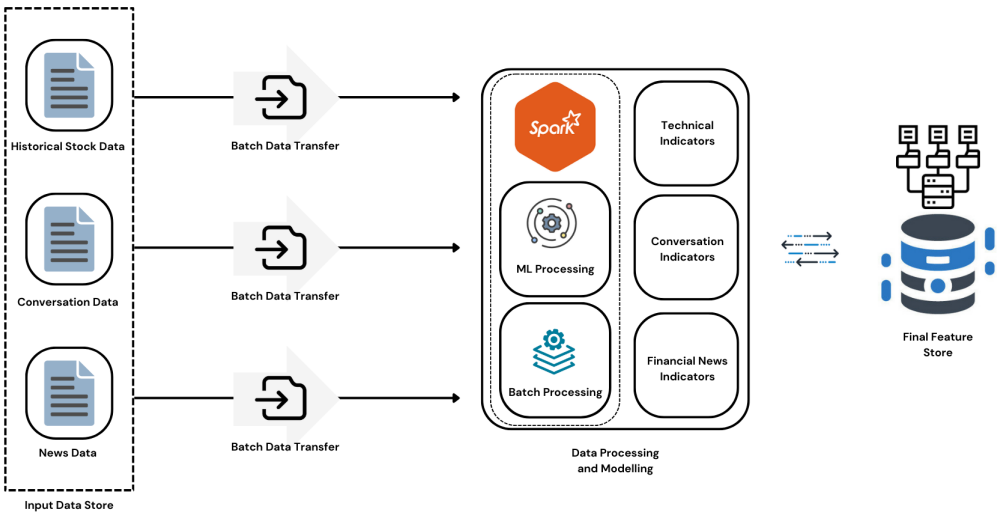

Figure 4: Architecture of Stock Movement Predictor

5.2 Tools and Technologies

- **Spark**: Spark is an open-source distributed general-purpose cluster-computing framework. Spark can be used to process and analyze data in a variety of different formats, including structured, semi-structured, and unstructured data. It is often used in conjunction with other tools, such as Apache Hadoop, to provide a complete big data processing solution.
- **HDFS**: HDFS, or the Hadoop Distributed File System, is a distributed file system that is designed to run on commodity hardware. It is the primary storage system used by Hadoop applications, and it provides a fault-tolerant and scalable platform for storing and processing large amounts of data. HDFS is designed to be highly scalable, with a single cluster capable of storing petabytes of data and supporting the simultaneous processing of millions of files.
- **Spark ML**: Spark ML is the machine learning library for Apache Spark. It provides a wide range of algorithms, tools, and APIs for building and deploying machine-learning models on Spark. The library includes popular algorithms for classification, regression, clustering, and collaborative filtering, as well as tools for feature engineering, model evaluation, and hyperparameter tuning.
- **Zeppelin**: Zeppelin is an open-source web-based notebook that allows users to create and share interactive data analytics. It is built on top of Apache Spark and provides a platform for data exploration, visualization, and collaboration. With Zeppelin, users can create and execute Spark queries and programs, as well as integrate them with other tools in the Hadoop ecosystem.

5.3 Flow

Now, let's try to understand the end-to-end flow of our architecture and how each sub-component comes into play.

- Input Data Store consists of data from all the data sources namely, Historical, Conversation and News stored as files on HDFS. HDFS ensures the high availability of data through replication.
- We ingest the data into our spark ecosystem by reading data as a batch in form of DataFrames. This helps both to keep data structured as well as allows Spark to use catalyst optimizer and catalog to apply optimizations.
- Then we design pipelines where each datasource is independently processed. In this step, we perform data cleaning, data profiling, and exploratory data analysis. We also try to understand what can be our labels for price movement prediction. So, this step helps us to build a hypothesis for empirical study.
- As a next step, we move to basic and advanced feature engineering to extract important features out of our data that can be useful for our hypothesis. For ex - We try to make features like Price, PriceVolume, and their relation with analysis. This showcases how important close price prediction is for price movement analysis.
- Finally, we apply data mining, machine learning, and advanced indicator engineering to get the indicators for our stock movement analysis for all the datasets. For ex: We calculate RSI, and Bollinger Bands that require window functions and analysis over time, we perform data mining and analytics on news and conversations data to find the correlation between price and indicators from these data.
- Now, the indicators gathered from all three datasets can be combined to get a unified view of a stock movement on a particular day.

6. Results

In this section, we will explore the results received after using the architecture and experiments discussed in the previous sections. We will try to deeply investigate the results and how they align with our experimentation.

6.1 Close Price Prediction:

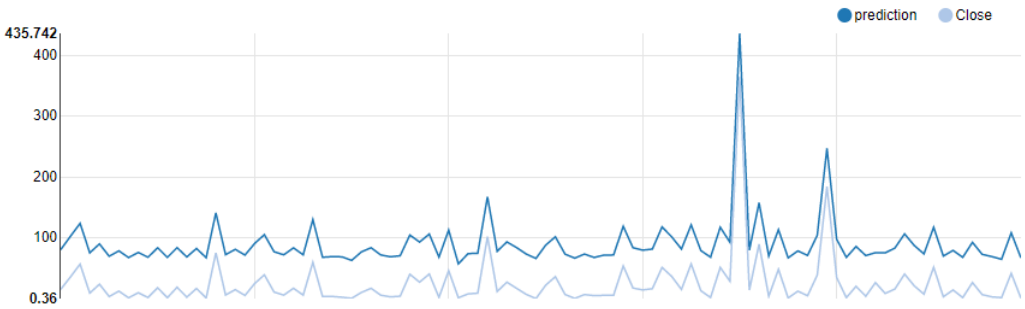

Figure 5: Prediction after Basic Feature Engineering

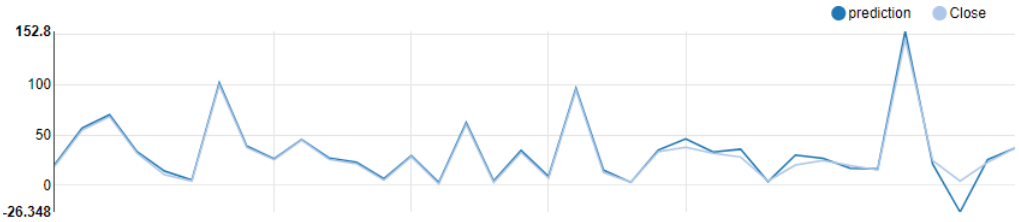

Figure 6: Prediction After Label Aware Feature Engineering

There are two figures attached above Fig 5 and Fig 6. Both figures show the distribution of the actual closing price against the predicted closing price.

The first figure i.e. Fig 5 shows the predictions from our machine learning model used to predict the closing price for a stock. It shows a huge deviation in the distribution of both the actual and predicted closing prices for the stock. Also, we calculated the RMSE and R2 to verify our calculations and we saw that the numbers for both losses were huge in their own respect. This proves that better feature engineering needs to be done.

The second figure i.e. Fig 6 also shows the predictions from our machine learning model used to predict the closing price for a stock. It shows very less deviation in the distribution of both the actual and predicted closing prices for the stock. This shows that our predictions closely follow the actual closing price for the stock. Also, we calculated the RMSE and R2 to verify our calculations and we saw that the numbers for both losses were very low in their own respect. This proves that the features engineered do a good job and this model can be used to predict the future closing price for the stock.

Finally, from the experimentation and the results we can say that we have enough data and features such that we can find means to predict closing prices for the stock which can be of great use to measure stock movement.

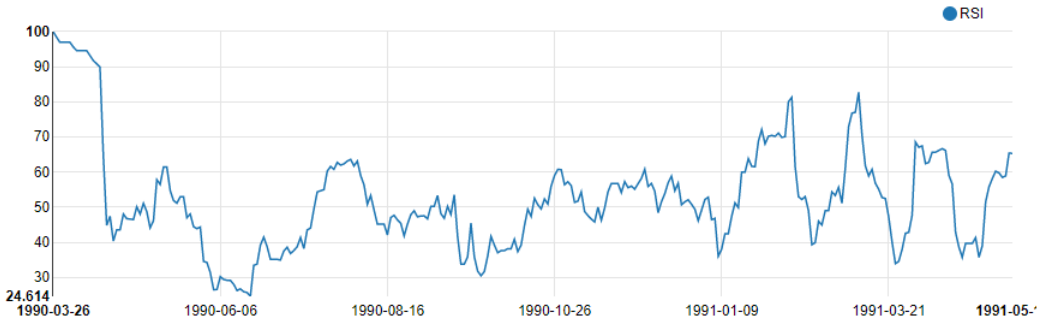

Figure 7: Prediction For Relative Strength Index

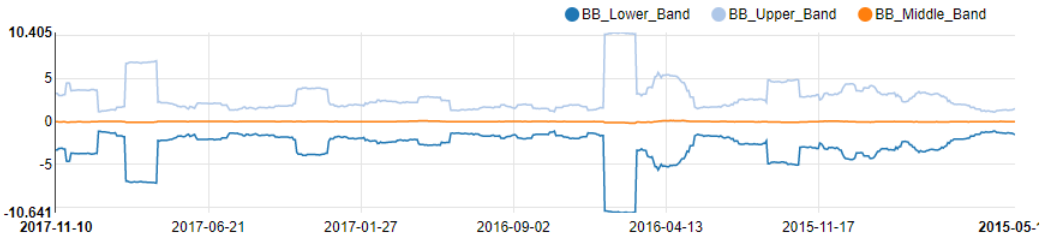

Figure 8: Prediction For Bollinger Bands

6.2 Technical Indicators Prediction:

There are two figures attached above Fig 7 and Fig 8. Both figures show the distribution of the two different technical indicators.

Relative Strength Index: The first figure i.e. Fig 7 represents RSI or Relative Strength Index. The relative strength index (RSI) is a technical analysis indicator used to measure the strength of a security's price action. It is calculated by taking the average of the gains and losses of a security over a given period of time, typically 14 days, and then comparing that average to the overall average price of the security. The resulting index is expressed as a number between 0 and 100, with high values indicating that the security is overbought and low values indicating that it is oversold.
Let's look at the mathematical expression for RSI calculation: Street (2022)

$$\textbf{RSI} = 100 - \left[\frac{100}{1 + \frac{\text{Avg Gain}}{\text{Avg Loss}}}\right] \tag{1}$$

Where,

$$\text{Init Avg Gain} = \left(\frac{\sum_{d=1}^{14} \text{Past Gain}}{14}\right) \quad \text{and} \quad \text{Init Avg Loss} = \left(\frac{\sum_{d=1}^{14} \text{Past Loss}}{14}\right) \tag{2}$$

$$\text{Avg Gain} = \left(\frac{[(\text{Avg. Gain} * 13) + \text{Cur Gain}]}{14}\right) \quad \text{and} \quad \text{Avg Loss} = \left(\frac{[(\text{Avg. Loss} * 13) + \text{Cur Loss}]}{14}\right) \tag{3}$$

Bollinger Bands: The second figure i.e. Fig 8 represents Bollinger Bands. Bollinger Bands are a type of technical analysis indicator that is used to measure the volatility of a stock price action.

They are calculated by taking the average price of a security over a given period of time, typically 20 days, and then plotting the standard deviation of that average above and below the average price to create an upper and lower band. The resulting bands are used to help traders identify potential entry and exit points in stock.
Let's look at the mathematical expression for Bollinger Bands calculation: Investopedia (2022)

$$\textbf{Upper Band} = MA(TP, n) + m * \sigma[TP, n] \tag{4}$$

$$\textbf{Lower Band} = MA(TP, n) - m * \sigma[TP, n] \tag{5}$$

$$TP = \frac{High + Low + Close}{3} \tag{6}$$

$$\text{Where, MA = Moving Average, TP = Typical Price} \tag{7}$$

$$\text{n = Number of days in the smoothing period (typically 20)} \tag{8}$$

$$\text{m = Number of standard deviations (typically 2)} \tag{9}$$

$$\sigma[TP, n] = \text{Standard Deviation over last n periods of TP} \tag{10}$$

Finally, from the experimentation, we can say that we have enough features to find different technical analysis indicators that can be used for price and stock movement tracking.

6.3 Correlation Between Public Conversation And Stock Movement:

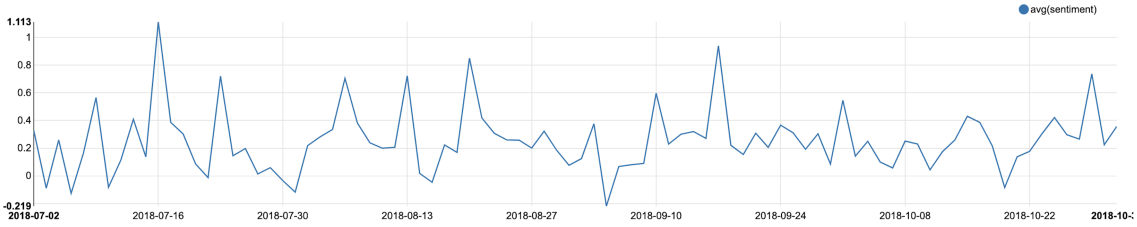

Figure 9: Average Sentiment per Day for TSLA

Figure 10: Price Movement per Day for TSLA

There are four figures attached above Fig 9-12. The first two figures present the relationship between average sentiment on Tesla and its corresponding price variations during July, 2018 to November, 2018. Similarly, the second pair illustrate the same relationship for Netflix on a different time frame.

Average Public Sentiment per Day: Figure 9 and Figure 11 represents the average sentiment score we generated from our public conversation dataset for Tesla and Netflix. We group the sentiment over stocks and calculate the average sentiment over all conversations that relate to a particular stock and present it in a time-series bar plot. The sentiment ranges between [-2, 2]. Positive and negative scores indicate positive and negative public attitudes correspondingly, and Zero represents neutrality.

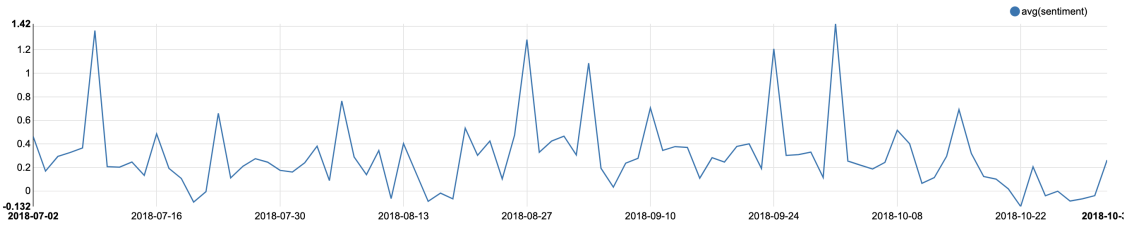

Figure 11: Average Sentiment per Day for NFLX

Figure 12: Price Movement per Day for NFLX

Price Movement per Day: Figure 10 and Figure 12 illustrate the price movement of the same stocks illustrated in Figure 9 and 11 within the same time frame. We obtain the data from Yahoo Finance by selecting the required stock names and the corresponding time frame.

Finally, from the experiment plots and results, we can notice that for Tesla, a sudden drop of sentiment at the beginning of August indicates the same drop of price at that time. Later sentiment ups and downs also reflect the price movement correspondingly. In the second example of Netflix, similar strong positive correlation can also be observed.

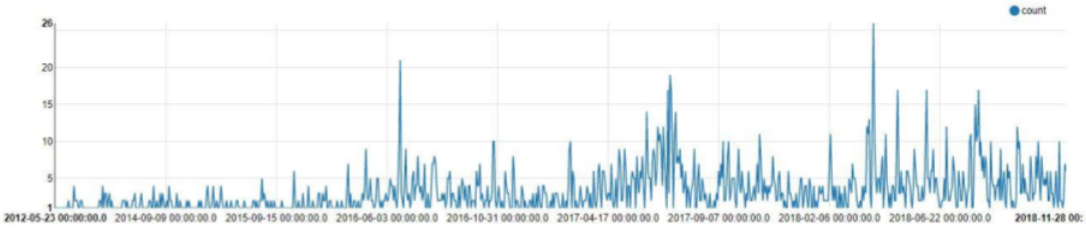

Figure 13: Number of News per Day for TSLA

Figure 14: Price Movement per Day for TSLA

6.4 Correlation Between News Conversation And Stock Movement:

Similarly, we presented four figures to illustrate the correlation between number of news mentioned for a stock with its price movement per day in a certain time frame. The first two figures illustrate the correlation on Tesla and the later ones on Netflix.

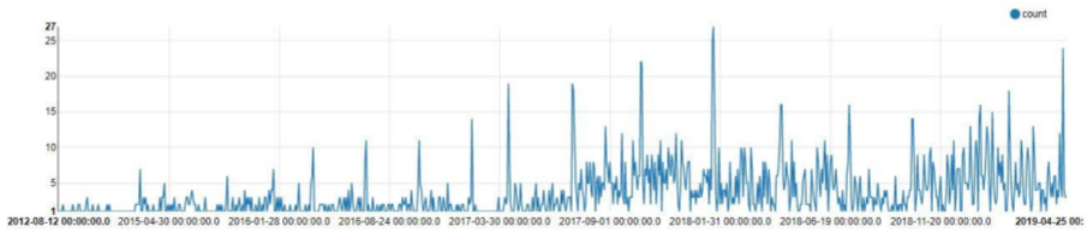

Figure 15: Number of News per Day for NFLX

Number of News per Day: Figure 13 and Figure 15 represents the number of news we generated from our public conversation data-set for Tesla and Netflix. We map news on weekends to weekdays as to accord with price data. Then we group the news over stocks and dates, calculate the number of news mentions for a particular stock and present it in a time-series plot as shown in those figures.

Figure 16: Price Movement per Day for NFLX

Price Movement per Day: Similarly as before, we obtain price data from Yahoo finance and present their variations in Figure 14 and Figure 16 for Tesla and NFLX correspondingly.

Finally, we observed that the news data analysis follows a normal distribution for the news frequency per day across stocks. Furthermore, from the previous plots, positive correlation between number of news and price movement can be detected. For example, we notice an increase of number of news for NFLX after Septemper, 2017, such increase can also be seen in the price movement.

7. Goodness of Hypothesis

- **Close Price Prediction**: We saw in the results section how we utilized exploratory data analysis, feature engineering, and machine learning, and modeling to predict the stock closing price for the day. We also saw that how the predicted value distribution closely follows the trend of actual value. Since this is our measure of correctness for the price prediction hypothesis, we can say that results and hypothesis are valid, which proves its goodness.
- **Technical Indicators Prediction**: For technical indicator engineering, we saw we engineered two technical indicators namely, Relative Strength Index and Bollinger Bands. For the Relative Strength Index, as per the domain knowledge, the RSI should remain between 30 and 70 and this is clearly visible from our graphs. This points in the direction that our results for RSI are correct. Also for Bollinger bands, we can say that they closely follow the general trend for upper, middle, and lower bands. We verified the calculations of mean and standard deviations of bands via hands and they seemed correct. This also points in to the direction of correctness. Now, also since both indicators are based on closing price, we need to ensure their correctness too. And we already proved it in our first hypothesis.
- **Correlation Between Public Conversation and Stock Movement**: We saw in the results section how we pre-processed our public conversation dataset, extracted the stock symbols and sentiments, grouped by stocks and dates and finally generated the average sentiment score for each stock. We also saw how the trend of price movement closely followed the trend of sentiment values. Since this is the measure of correctness for the correlation between public conversation and price movement hypothesis, we can say as the results illustrated, the hypothesis we previously made is valid, which proves its goodness.
- **Correlation Between News and Stock Movement**: We saw in the results section how we pre-processed our public news dataset, grouped news by stocks and dates and finally generated the number of news for each stock. We also saw how the trend of price movement generally

aligned with the trend of number of news per day. Since this is the measure of correctness for the correlation between public news and price movement hypothesis, we can say as the results illustrated, the hypothesis we previously made is valid, which proves its goodness.

8. Challenges

The historical stock dataset has time-series properties and requires advanced feature engineering. This involved the use of lags, windows, and aggregations over windows using UDF's. Deriving features from the news dataset cannot be directly achieved by the sentiment analyzer due to the large size of news content per stock. There is no support for deep learning model training which proposes challenges for sentiment analysis and related tasks. News data is biased by the publisher. The conversation dataset involves multiple tickers and hence requires some additional preprocessing to fetch related tickers. The conversation comments also include non-standard characters which are difficult to handle.

9. Obstacles

For the Historical Stock Data, one of the biggest challenges was technical indicators engineering. The complexity of handling time-series data makes it a difficult coding problem. For the Conservation Data, one of the biggest challenges was handling the non-standardization of the conversation column. Applying symbol extraction through regex is a bit challenging. For the Financial News Data, one of the biggest challenges was handling the scale of the news column. There is no limit on the length of the news column and hence requires additional preprocessing steps to get a summarized view.

10. Future Scope

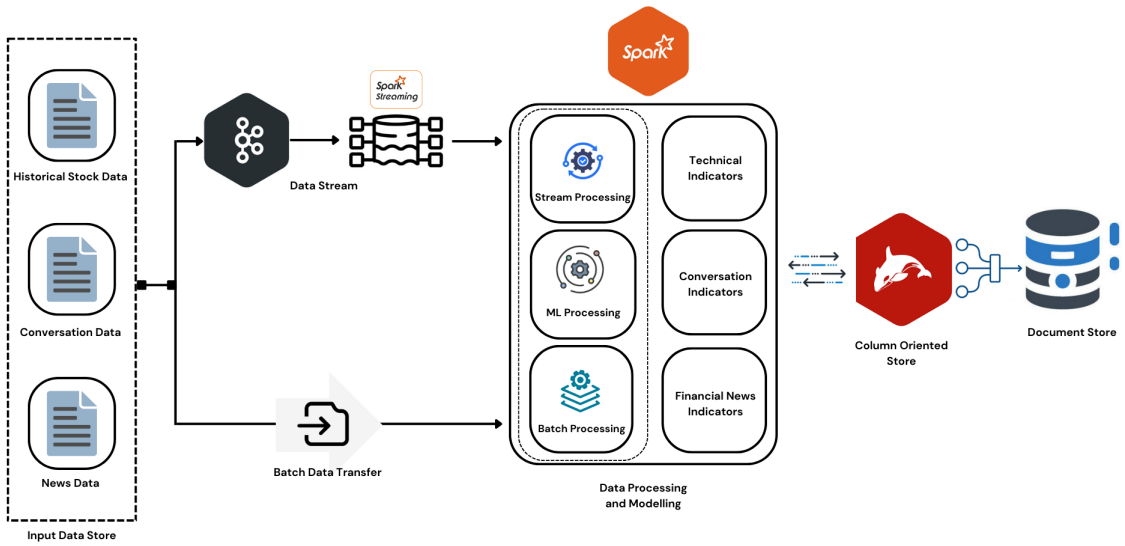

Figure 17: Future Scope

There is a scope for improvements in the discussed architecture and in this section, we discuss possible improvements.

- Introducing Kafka and Spark Streaming will make it possible to make predictions in near real-time.
- Introducing Column Oriented Store like HBase will make it possible to store intermediary features e.g having three-column families for each of the data sources.
- Having a final document store for the unified and comprehensive document of the stock indicators that are to be displayed on the dashboard.

From an analysis view, there is a lot of scope for performing advanced technical indicator engineering. There are a lot of technical indicators which can be used collectively for price movement prediction. Since this is version 1.0, we worked more on understanding the possibility of engineering such features.

11. Summary

This project introduces how one can design a system that can help us to profile stock market data and its movement. Designing a system to profile stock market data and its movement involves analyzing and organizing large amounts of data about publicly traded companies and the stock market as a whole. This can include data about a company's financial performance, market trends, and other factors that may impact the value of its stock. By using various data sources and analyzing this data in a systematic way, it is possible to gain insight into a company's performance and make informed predictions about the movement of its stock.

This project also shows how this can be achieved by using heterogeneous data sources, which refers to using multiple types of data from various sources. This can include financial data from a company's financial statements, news articles about the company, social media activity related to the company, and other types of data. By combining and analyzing this diverse range of data, it is possible to gain a more comprehensive understanding of a company and its performance.

Using the output of this system, it is possible to build powerful tools and dashboards that can help to solve challenging engineering problems related to the stock market. These tools could be used to track stock prices, analyze market trends, and make predictions about the performance of different companies. They could also be used to identify opportunities for investment or to help inform trading decisions. By building these tools and dashboards, it is possible to gain a deeper understanding of the stock market and make informed decisions about investing in it.

Acknowledgments

We would like to thank Prof. Yang Tang for helping us to understand the Big Data Infrastructure, Design, and Engineering in the lectures and by guiding us post lectures and over mail to resolve the roadblocks. We would also like to thank Team HPC for providing us with the wonderful platform and support that made this project possible. We would also like to thank the contributors of the respective data sources for their wonderful contribution to the open source community. This project was carried out as a part of the curriculum for the Big Data Application Development course at NYU Courant Institute of Mathematical Science.

References

Investopedia. Bollinger bands: What they are, and what they tell investors, 2022. URL https://www.investopedia.com/terms/b/bollingerbands.asp.

Boris Marjanovic. Historical daily prices and volumes of all u.s. stocks and etfs, 2017. URL https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350, 2016. doi: 10.1109/SCOPES.2016.7955659.

Mostafa Mohamed Seif, Essam M Ramzy Hamed, and Abd El Fatah Abdel Ghfar Hegazy. Stock market real time recommender model using apache spark framework. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 671–683. Springer, 2018.

The Street. What is the relative strength index? definition, calculation example, 2022. URL https://www.thestreet.com/dictionary/r/relative-strength-index-rsi.

Linda L. Tesar and Ingrid M. Werner. U.S. Equity Investment in Emerging Stock Markets. volume 9, pages 109–129, 01 1995. doi: 10.1093/wber/9.1.109. URL https://doi.org/10.1093/wber/9.1.109.

Zhaoxia Wang, Seng-Beng Ho, and Zhiping Lin. Stock market prediction analysis by incorporating social and news opinion and sentiment. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1375–1380, 2018. doi: 10.1109/ICDMW.2018.00195.