Social Media based Stock Market Analysis using Big-Data Infrastructure

Group ID: G2_10

Shashanka Venkatesh
Venkataraman Nagarajan
Vishakan Subramanian

SSN College of Engineering, Chennai

March 31, 2022

Outline

Abstract

Observation

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

Analysis

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. In our proposed methodology, we consider Twitter, a popularly used micro-blogging site as our social media source. We perform sentiment analysis on collected Tweets, and measure their correlation to stocks of companies within the sector the Tweet appeals to.

Architectural Design

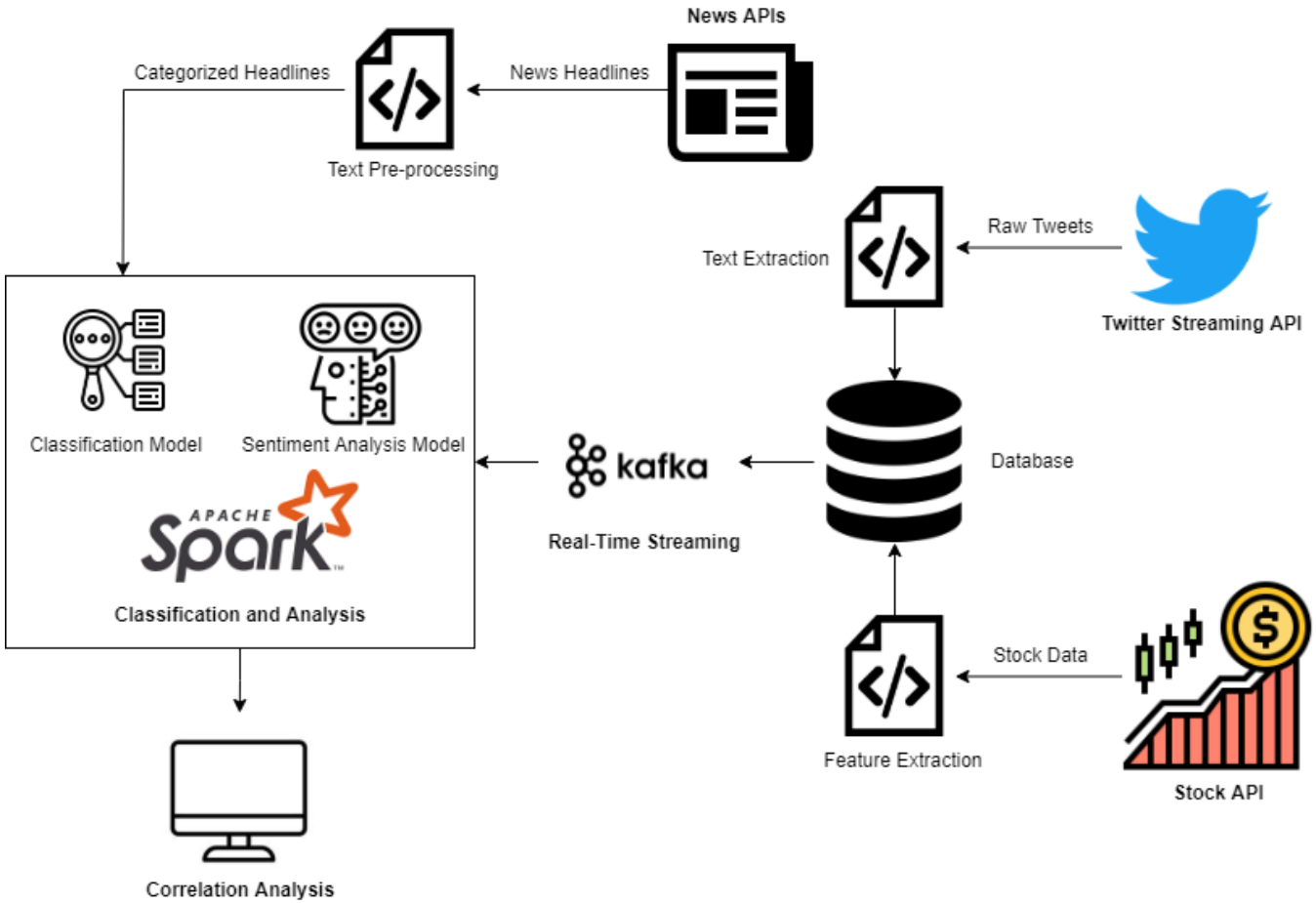

Figure: Proposed Architecture

Architectural Design (Category Classifier Model)

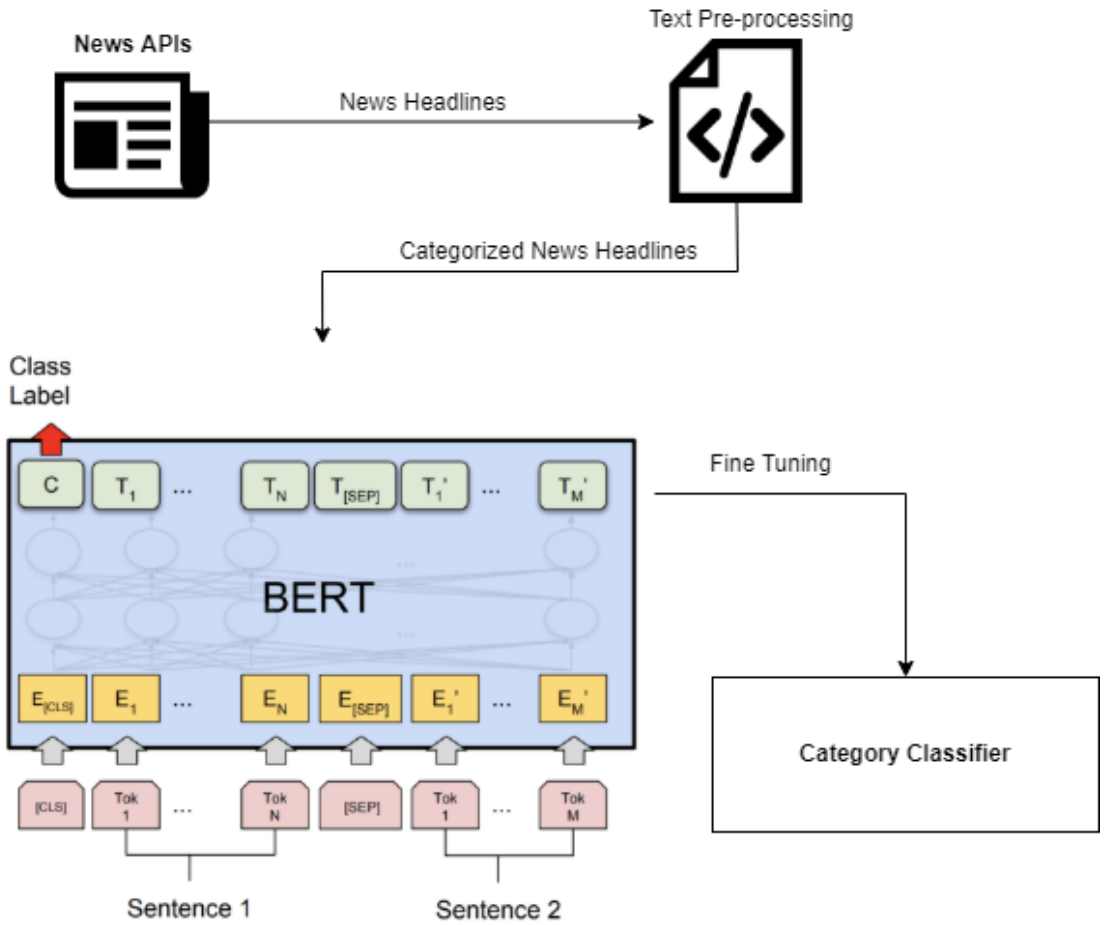

Figure: Fine tuning the BERT model using News Data

Architectural Design (Data Processing)

- Twitter Data Processing
  - The Twitter Streaming API is used to collect tweets containing given keywords/hashtags.
  - The tweets are then stored in a SQLite database.
- Stock Data Processing
  - The Polygon API is used to collect stock ticker data.
  - historic Stock data for the past two years, and was stored in a dataset categorized according to the sector of the companies queried.
  - The data is then stored in the SQLite database.

Architectural Design (Correlation Analysis)

- Tweets and Stock market ticker data is streamed from the SQLite database using Apache Kafka to Apache Spark.
- The category classifier is used to remove false-positive tweets, and their sentiment is analyzed with the BERTweet sentiment analyser.
- The correlation between the aggregate ticker data of companies belonging to a sector, and the aggregate sentiment of tweets about the same sector is analysed using the Pearson's coefficient of correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Figure: The Pearson Coefficient for a sample

Justification of Problem Statement

The Essential Reason

Presently, with the majority of the world being online, people tend to be more engaged on social media & use it a a source of news. Market enthusiasts are increasingly considering the opinions reflected by social media. Finding the correlation between such events & stock valuations is hence valuable.

Using the BERT Model

- From our literature survey, we understood that several NLP techniques like n-grams[7], TF-IDF (for Feature Vector generation) along with Random Forest and Naive Bayes classification techniques[1] were used.
- BERT has pushed the state-of-the-art for 11 NLP tasks[10] and has proven to be revolutionary for NLP tasks. Since we have a text classification & sentiment analysis task, we considered using BERT - for its modern ubiquity and its high results in the field of NLP.
- Emphasizing the above point, we use BERTweet[12], a model trained on 850M English Tweets. It possesses the same architecture as BERT base and uses the pre-training procedure of RoBERTa, and outperforms competitor NLP models on the SemEval 2017 Task 4A (Twitter Sentiment Analysis)

Using Apache Kafka and Apache Spark

- We use Apache Kafka to streamline & homogenize the data ingestion process to the Spark architecture where the classification & analysis of data is performed. To simulate real-time event processing with reliability and low latency, Kafka is an obvious candidate for the job.[4]
- Spark is a popular tool in the scenario of real-time data analytics. Since Tweet data and stock market data are voluminous with high velocity, using traditional batch processing has its limitations. Stream processing can be utilized to update the average sentiment of Tweets for a particular sector in real-time, as they are ingested into the system.
- The map, reduce & window operations help us aggregate data consistently. MLlib is used to perform correlation analysis and provide visualizations of the output.

Techniques Used

- News Data Collection APIs
  - NewsData
  - NewsCatcher
  - New York Times
- The Twitter API along with keywords for each category of analysis was used to collect tweets.
- The Polygon API is used to collect stock ticker data.
- Python script were used to collect, pre-process and store the data in CSV-format datasets.
- The datasets are then stored in an SQLite Database using another Python script.

Techniques Used - Results

- The news-based Tweet classification model developed by Lee et al.[1] using Naive-Bayes method was trained on 5,000 datapoints & yielded an accuracy of 77%.
- Comparatively, our classifier for the same task was trained on 10,000 data points using bert-base-uncased & yielded an accuracy of 96%.
- Tweet sentiment analysis performed by Kanavos et al.[7] using Stanford Core NLP and Naive Bayes techniques using bigrams yielded an accuracy of nearly 65%. (156K Tweets)
- Comparatively, the BERTweet sentiment analyzer has reported an accuracy of 72% over a much larger dataset.[12] (850M Tweets)

Progress since previous review

Data Collected from News APIs

	A	
1	Crypto	What's Driving Ethereum, Dogecoin, and THORChain Higher Today?
2	Crypto	The cryptocurrency boom has spawned enterprises democratically governed by a community of users. Or that's the theory. Making it work has been much messier.
3	Oil	Share Market Live: SGX Nifty Gains 1.7%; ITC, Zomato, Punjab National Bank, Paytm, United Spirits In Focus
4	EVs	The popularity of battery-powered cars is soaring while the overall auto market stagnates, a worldwide trend.
5	Oil	A nuclear threat, and skepticism at potential talks.
6	Gaming	Games Inbox: GTA 5 triple dipping, XCOM 3 hopes, and Stranger Of Paradise sales advice
7	Gaming	The deal follows two other major acquisitions in the game market by Microsoft and Take-Two Interactive.
8	Crypto	Why Bitcoins Energy Problem Is So Hard to Fix: QuickTake
9	Gaming	PS to stream 'Hogwarts Legacy' State of Play on March 17
10	Oil	Shipping companies, which can command more than $200,000 a day for a vessel, are cashing in on increased demand.
11	Gaming	A new video game will let players pretend to be the politicians they most admire — or despise.
12	Tech	In this regressive tale of demons and damsels, a priest must admit his sins before he can vanquish a malevolent spirit.
13	Oil	Oil Price Benchmarks Fall Below $100, First Time In Weeks
14	Crypto	Katie Haun and I discuss the future cryptocurrency could create.
15	Gaming	Xbox Game Pass Ultimate Adds Touch Controls for Among Us and Eight More Games
16	EVs	Ready for an electric vehicle future? The Volkswagen ID.4 may be the electric vehicle that carries you into it.

Figure: Sample News Data obtained from the News APIs

Progress since previous review (cont.)

Data Collected from Twitter API

	A	B	C
55	Crypto	2022-03-23	I wonder why Chinese people can buy #BTC from exchange even they stay in China mainland my friends told me they can receive #bitcoin from customer also,that mean Chinese government can not ban #bitcoin
56	Gaming	2022-03-23	A person who never made a mistake never tried anything new. So try my #EA #FX #ROBOT and taste your winning moment everyday. PM for details
57	EVs	2022-03-23	@thatstocksgirl Investing in #TSLA now is like investing in #amazon back in 2010
58	EVs	2022-03-23	#TSLA This is called manipulation kids.
59	Crypto	2022-03-23	@11MinutesNFT This will certainly be a successful project as it has a great team that is more than qualified and focused on making this project a success. #11MINUTES #BTC #CRYPTO
60	Crypto	2022-03-23	Give a man a gun and he can rob the bank. Give a man a bank and he can rob the world #Hyperinflation #BTC #cryptocurrencies #Banks
61	EVs	2022-03-23	Our longer term swing trading portfolio now consists of buy signals in #ES, #NASDAQ, #DXY, #BTC, #Gold, #CrudeOil and #TSLA and a sell signal for #Bonds
62	Oil	2022-03-23	Closed Buy 6.71 Lots of #COP.N 101.0 for +3.0 pips, total for today +625.9 pips
63	EVs	2022-03-23	@cinemaheist @DaveHcontrarian Imagine a balloon with a puncture. You can blow air into it, it may even grow bigger, but it is constantly deflating, and you'll run out of air. When rates rose to 1.5 last year the bubble burst. The story is set in stone for #TSLA.
64	Crypto	2022-03-23	After this close above $42,600 #BTC send everything! Told y"all I expect a pump tonight. #SUSHI. #COMP #VET #FIL #XRP just send everything
65	EVs	2022-03-23	#TSLA binging over $1000 again. Woot!!
66	EVs	2022-03-23	@elonmusk Adding #bitcoin to #TSLA balance sheet was a genius idea
67	Crypto	2022-03-23	4291) Current #Bitcoin Price is $42539 #BTC #Crypto Indicators Daily: -RSI: 57.2 -MA(20): 40130 -MA(50): 40706 -MA(200): 48351 -Bollinger B. lower/upper: 37077/43184 Current #Ethereum Price is $3011 #ETH -RSI: 61.0 -MA(20): 2715 -MA(50): 2811 -MA(200
68	EVs	2022-03-23	Forgot to add #TSLA
69	Crypto	2022-03-23	The latest #bitcoin block 728722 with 739 transactions was just mined by Foundry USA Total Fees:0.00730075 Block Subsidy:6.25 #Bitcoin #BTC #blockchain #analysis
70	EVs	2022-03-23	@vtnicholas Thank you. Doesn't always work out that well but #TSLA has been good lately.
71	Gaming	2022-03-23	By failing to prepare, you are preparing to fail. However By purchasing #EA #FX ROBOT, you are preparing to #Win in #Forex. PM for Details!
72	Crypto	2022-03-23	@SoccerFan_io big project #Ethereum #LINK #eth
73	Crypto	2022-03-23	@mcshane_writes @nayibbukele Those whom will struggle the most in the New Economy, will be those who try to drag their "old money" paradigms around with them.. #Crypto #bitcoin #btc #dogecoin #Ethereum
74	EVs	2022-03-23	@iluvstocks1 the most amazing room. With the most amazing Vegas. If you want to learn the stock market Vegas is the person to go to. Thank you for an awesome green day. #TSLA #BABA #AMC #MARA
75	EVs	2022-03-23	Wednesday, March 23, 2022 Daily Market Correction : Waning 72% in : 0705 - 0908 EST : 1009 - 1211 EST : 1414 - 1616 EST : 1718 - 1920 EST : 2117 - 2314 EST $SPY #sp500 #Bitcoin $AAPL $NVDA $MSFT #TSLA #Googl #BA (not financial advice)
76	Oil	2022-03-23	we often hear the word COP. why the word cop represent policeman? what is the association? #cop #Police #English #KnowMore #do_you_know?

Figure: Tweets collected from Twitter API based on Ticker hashtags

Progress since previous review (cont.)

Data Collected from Stock API

	A	B	C	D	E	F	G	H	I
1	CATEGORY	TICKER	DATE	OPEN	CLOSE	PRE_MARKET	AFTER_HOURS	HIGH	LOW
1504	Gaming	EA	2022-03-21	125.09	125.56	125.04	125.56	126.4	124.6377
1505	Tech	INTC	2020-03-30	52.99	55.49	52.23	55.4	56.1	52.83
1506	Tech	FB	2020-03-30	159.18	165.95	155.31	166.29	166.75	158.06
1507	EVS	TSLA	2020-03-30	102.052	100.426	101	100.6	103.3294	98.246
1508	Oil	CVX	2020-03-30	68.01	71.95	68.1	71.75	72.93	67.02
1509	Oil	XOM	2020-03-30	36.23	37.5	36.95	37.45	38.15	35.86
1510	Oil	COP	2020-03-30	28.51	29.29	29	29.42	30.14	27.52
1511	Tech	INTC	2020-03-31	55.06	54.12	55.38	53.4	55.795	53.22
1512	Tech	FB	2020-03-31	165.48	166.8	166.35	164.75	170.93	164.2
1513	EVS	TSLA	2020-03-31	100.25	104.8	101.8	103.1	108.592	99.4
1514	Oil	CVX	2020-03-31	73.27	72.46	74.01	71	75.04	70.81
1515	Oil	XOM	2020-03-31	38.34	37.97	38.25	37.24	39.45	37.37
1516	Oil	COP	2020-03-31	30.03	30.8	31	29.81	31.92	29.71
1517	Tech	INTC	2020-04-01	52.5	51.88	52.92	52.11	54.69	51.43

Figure: Sample Stock Data

	A	B	C	D	E
1	CATEGORY	TICKER	DATE	OPEN	CLOSE
2	Crypto	ETH-USD	2020-03-30	124.43	132.32
3	Crypto	BTC-USD	2020-03-30	5878.98	6406.4
4	Crypto	ETH-USD	2020-03-31	132.32	132.7
5	Crypto	BTC-USD	2020-03-31	6406.4	6421.42
6	Crypto	ETH-USD	2020-04-01	132.701	136.45
7	Crypto	BTC-USD	2020-04-01	6424.02	6666.11
8	Crypto	ETH-USD	2020-04-02	136.14	141.8
9	Crypto	BTC-USD	2020-04-02	6666.11	6804.52
10	Crypto	ETH-USD	2020-04-03	141.177	141.327

Figure: Sample Crypto Stock Data

Progress since previous review (cont.)

Category Classifier Model Summary

	precision	recall	f1-score	support
0	0.95	0.95	0.95	268
1	0.98	0.97	0.97	390
2	0.96	0.96	0.96	533
3	0.95	0.94	0.95	284
4	0.96	0.97	0.97	626
accuracy			0.96	2101
macro avg	0.96	0.96	0.96	2101
weighted avg	0.96	0.96	0.96	2101

Figure: Category Classification Summary

A sample of approx. 2000 Tweets were used as test data.

Progress since previous review (cont.)

Sentiment Analysis Model

BERTweet is used for sentiment analysis of the tweets, it has been proven to give an accuracy of 72% with a F1 score of 72.5%

```
[17] sentiment_score("Looks like the EVs are the next big thing")

     1

[18] sentiment_score("Something is horribly wrong with the crypto market!!! Do not invest!")

     -1

[19] sentiment_score("Extensions: someone can add a filter in Twitter to help users get relevant tweets")

     0

[21] sentiment_score("Plugin ideas: you should make a web browser plugin to convert any #fiat to #bitcoin \
     For example: Instead of this costs $50 a night, it should change it into 0.000192 BTC a night")

     0
```

Figure: Sample output from BERTweet Model

Discussion

Correlation Analysis Between tweet sentiments and %Δ stock price

1. Let x_1 be the average sentiment of the observed tweets on a given day d, x_2 be the average sentiment of the observed tweets on day $d-1$ and y be the %Δ stock price on d. Consider two approaches to find a correlation between tweet sentiments and %Δ stock price
2. Approach 1: Assuming the tweets affect the stock market immediately
   - Lets try to make a table of values x_1 and y for d for D given days
   - Now with a table, we try to find whether there is a correlation between the two values on the table using Pearson Coefficient correlation.
3. Approach 2: Assuming the tweets take time to have an effect on the stock market
   - Similar to Approach 1, but with a table of x_2 and y

Expected Outcomes

The main purpose of the entire analysis, is to unearth a correlation between Tweets and Stock market changes, if any. The expected outcome, therefore, is an accurate correlation analysis between tweet sentiments, and the stock market variations in the related sectors.

Further Work

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

Timeline

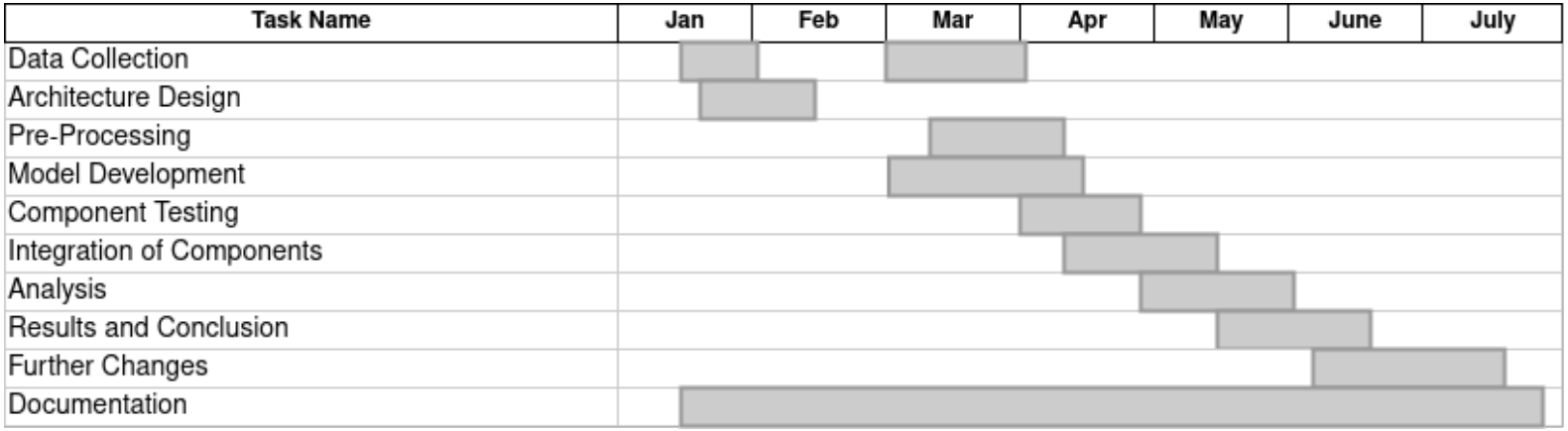

Figure: Timeline Chart

References


Chungho Lee, Incheon Paik, *Stock Market Analysis from Twitter and News Based on Streaming 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)


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)


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


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


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

References (cont.)


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


Marsland, Stephen, *Machine Learning: An Algorithmic Perspective*, 2014


Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *BERT: Pre-training 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*

Attendance Sheet

ATTENDANCE REPORT FOR FINAL YEAR PROJECT 2021-2022: REVIEW 1

DATE	WORK DONE DURING THE WEEK	SUPERVISOR SIGNATURE
07 March 2022	Data collection is being done. Twitter API Essential Access is granted and Tweet data based on hashtags are fetched and saved in local system.	
14 March 2022	A machine learning model to categorize news data fetched from the APIs is being built and the work to build a model to obtain Tweet sentiments is started.	
21 March 2022	Stock data is being fetched from Polygon API. The Twitter sentiment analysis is completed. Script is written to migrate data from CSV files to a database environment. PySpark is setup and configured for streaming the data from database environment for further correlation analysis.	

Thank You

Thank You
Q & A