

# Social Media based Stock Market Analysis using Big-Data Infrastructure

Group ID: G2 10

Shashanka Venkatesh  
Venkataraman Nagarajan  
Vishakan Subramanian

Mentor: Dr. N. Sujaudeen

SSN College of Engineering, Chennai

April 22, 2022

# Outline

Stock Market  
Analysis

Shashanka,  
Venkataraman  
& Vishakan

Project  
Objectives

Architectural  
Design

Completed  
Outcomes

Expected  
Outcomes

References

- 1 Project Objectives
- 2 Architectural Design
- 3 Completed Outcomes
- 4 Expected Outcomes
- 5 References

# Project Objectives

Stock Market  
Analysis

Shashanka,  
Venkataraman  
& Vishakan

Project  
Objectives

Architectural  
Design

Completed  
Outcomes

Expected  
Outcomes

References

## 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 objective of this research is to analyse the effect of social media trends over the value of a stock at any given time. To capture public mood, we use Twitter as our social media entity. We perform sentiment analysis on collected Tweets, and measure their correlation to stocks of relevant market sectors.

# Architectural Design

## Stock Market Analysis

Shashanka,  
Venkataraman  
& Vishakan

## Project Objectives

## Architectural Design

## Completed Outcomes

## Expected Outcomes

## References

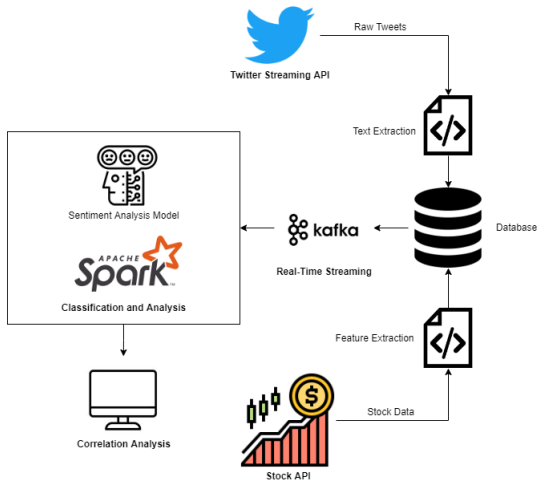


Figure: Proposed Architecture

# Architectural Design (Data Processing and Correlation Analysis)

Stock Market  
Analysis

Shashanka,  
Venkataraman  
& Vishakan

Project  
Objectives

Architectural  
Design

Completed  
Outcomes

Expected  
Outcomes

References

- The Twitter Streaming API is used to collect tweets containing given keywords/hashtags. The Polygon API is used to collect historic stock data for the past two years.
- Tweets and Stock market ticker data is streamed from the SQLite database using Apache Kafka to Apache Spark.
- Tweets are analyzed for their sentiment with the BERTweet sentiment analyzer.
- 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  $\bar{y}$

# Results of Completed Modules

## Stock Market Analysis

Shashanka,  
Venkataraman  
& Vishakan

## Project Objectives

## Architectural Design

## Completed Outcomes

## Expected Outcomes

## References

category	tweetDate	count	tweet
Search column...	Search column...	Search column...	Search column...
1 Tech	2020-04-07	10	BLOCK TRADE detected in #GOOGL
2 Tech	2020-04-07	10	@PlayAdoptMe I like it #goog
3 Tech	2020-04-07	1	Apple Inc price at close, 2020-04-07, is 271.1. #apple #AAPL
4 Tech	2020-04-07	1	Apple Inc stock rose by 3.288%! Currently priced at 271.1. #apple #AAPL
5 Tech	2020-04-07	1	Looking into the chart of #AAPL makes you wonder if it was just a correction and not a crash. That strong, respected key level perfectly

Figure: Tweets collected from Twitter API based on Ticker hashtags

category	ticker	stockD...	open	close	pre_market	afterHours	high	low
Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...
1 Tech	GOOGL	2020-03-26	1114.72	1162.92	1090.32	1163	1171.48	1092.03
2 Tech	AAPL	2020-03-26	61.63	64.61	60.8	64.7325	64.67	61.59
3 Gaming	EA	2020-03-26	90.47	99.2	88.38	99.23	99.57	90.08
4 Tech	GOOGL	2020-03-27	1127.47	1110.26	1123.91	1109.79	1151.05	1104.0027
5 Tech	AAPL	2020-03-27	63.1875	61.935	63.61	61.75	63.9675	61.7625

Figure: Stock data collected from Polygon API based on tickers

# Results of Completed Modules (cont.)

## Stock Market Analysis

Shashanka,  
Venkataraman  
& Vishakan

## Project Objectives

## Architectural Design

## Completed Outcomes

## Expected Outcomes

## References













	category   	date   	score   	count   
	Search column...	Search column...	Search column...	Search column...
1	Tech	2020-04-13	32	453
2	Tech	2020-04-14	192	585
3	Tech	2020-04-08	-51	416
4	Tech	2020-04-12	6	210
5	Tech	2020-04-07	149	217

Figure: Aggregated Tweet Data










	category   	date   	agg_percent   
	Search column...	Search column...	Search column...
1	Tech	2020-04-16	0.0045
2	Tech	2020-04-17	-0.88175
3	Tech	2020-05-01	-0.37625
4	Tech	2020-05-12	-2.2245
5	Tech	2020-06-22	1.205

Figure: Aggregated stock Data

# Results of Completed Modules (cont.)

Stock Market  
Analysis

Shashanka,  
Venkataraman  
& Vishakan

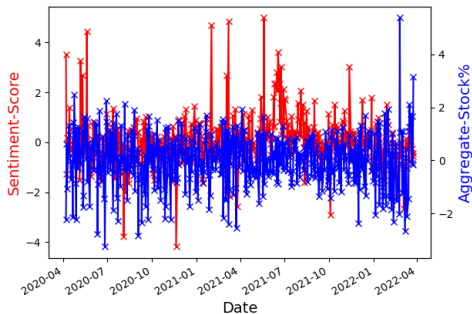
Project  
Objectives

Architectural  
Design

Completed  
Outcomes

Expected  
Outcomes

References



**Figure:** Tech - Time Series plot between Sentiment Score and Stock Price Change



# Expected Outcomes

## Stock Market Analysis

Shashanka,  
Venkataraman  
& Vishakan

## Project Objectives

## Architectural Design

## Completed Outcomes

## Expected Outcomes

## References

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.

If a significant correlation is found between a market sector and its respective Tweets, it is worthwhile to predict the market's performance using the observed social media trends. Techniques like regression could be used for prediction.

## References

## Stock Market Analysis

## 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<sup>nd</sup> 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*



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



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*

# Thank You

Stock Market  
Analysis

Shashanka,  
Venkataraman  
& Vishakan

Project  
Objectives

Architectural  
Design

Completed  
Outcomes

Expected  
Outcomes

References

**Thank You**  
**Q & A**