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

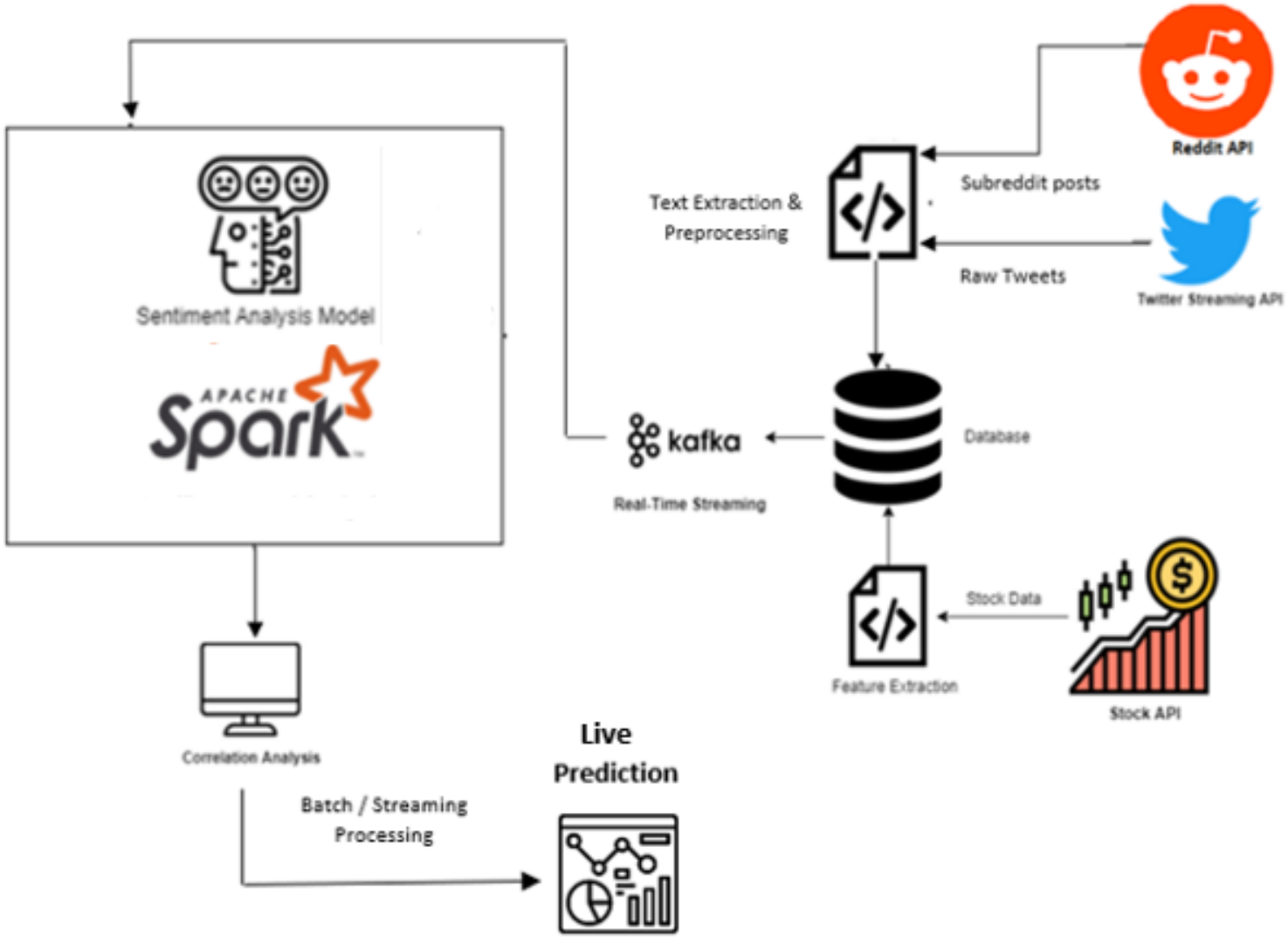

Figure: Proposed architecture of the system

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.

Results of the Completed Modules

	c...	d..	title	count
	Search colum	Search colur	Search column...	Search column...
1	CATEGORY	DATE	TITLE	COUNT
2	Health	2022-08-01	@Tyler_M_Poe @SenWhitehouse Hmh. Reminds me. Weed killer was in the worx. Paraguat, analog to agent ora...	1
3	Health	2022-08-01	Future Wolf Springs Tooth. I. "Unh "I Nabe Pa NTR [Sad News] It Next Door Beautiful Girls Forgotten And ALL R...	1
4	Health	2022-08-01	I guess Im going to this Black Alumni thing I never participate in anything UNH related since I graduated	1
5	Health	2022-08-01	$UNH hits $539, this will fly to $543 its nice price resistance. It's current price is over the 20 day SMA; making it ...	1
6	Health	2022-08-01	@Tyler_M_Poe Stop. You must know, in our DNA. Epigenome means...if you edit 3gacd. Unh no, nooo	1
7	Health	2022-08-01	We have four stocks on our swing watchlist for Tuesday 8/2. Tweets coming shortly on $VZ, $UNH, $BDX, and $...	1
8	Health	2022-08-01	@thevpo1 @DataProgress Do you have a link to the UNH poll?	1
9	Health	2022-08-01	How the largest stocks performed today Apple $AAPL -0.6% Microsoft $MSFT -1% Google $GOOGL -1% Amaz...	15
10	Health	2022-08-01	tommorow is also a day unh	1
11	Health	2022-08-01	Top 5 best and worst performers from the S & ;P 500 today 8/1/2022 #SP500 Best: 1. $BA: +6.22% 2. $PG: +2.8...	1
12	Health	2022-08-01	@LaraHReid Looked it up. Still there! Cedar Waters in Nottingham, near Epping. I used to drive by the sign on ...	1
13	Health	2022-08-01	@JacksonDowney11 @Coach_Borden @rwsantos2 @Coach_DeAndrade @UNH_Football Worked for this! Cong...	1

Figure: Health Tweet Reddit data - Sample from DB

	category	ticker	stockDate	open	close	high	low
	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...
1	Health	UNH	2022-08-01	542.27	535.38	543.65	531
2	Health	UNH	2022-08-02	537.55	535.46	540.53	531.72
3	Health	UNH	2022-08-03	537.26	540.65	543.62	535.69
4	Health	UNH	2022-08-04	539.29	533.75	540.43	532.755
5	Health	UNH	2022-08-05	530	535.06	535.5	527.56
6	Health	UNH	2022-08-06	534.95	535.83	538.4	530.49
7	Health	UNH	2022-08-07	537.43	536.22	539.85	531.95
8	Health	UNH	2022-08-08	539.91	536.6	541.3	533.42
9	Health	UNH	2022-08-09	540	537.26	545.74	536.38
10	Health	UNH	2022-08-10	540.26	537.72	541.98	533.81
11	Health	UNH	2022-08-11	538.24	532.86	538.6336	529.675
12	Health	UNH	2022-08-12	535.61	543.7	544.395	534.45
13	Health	UNH	2022-08-13	538.64	544.17	545.8	536.5

Figure: Health Stock (UNH) from Db

Results - Contd.

Partition Information

Partition	Latest Offset	Leader	Replicas	In Sync Replicas	Preferred Leader?	Under Replicated?
0	60,440	0	(0)	(0)	true	false

Figure: Kafka partition information

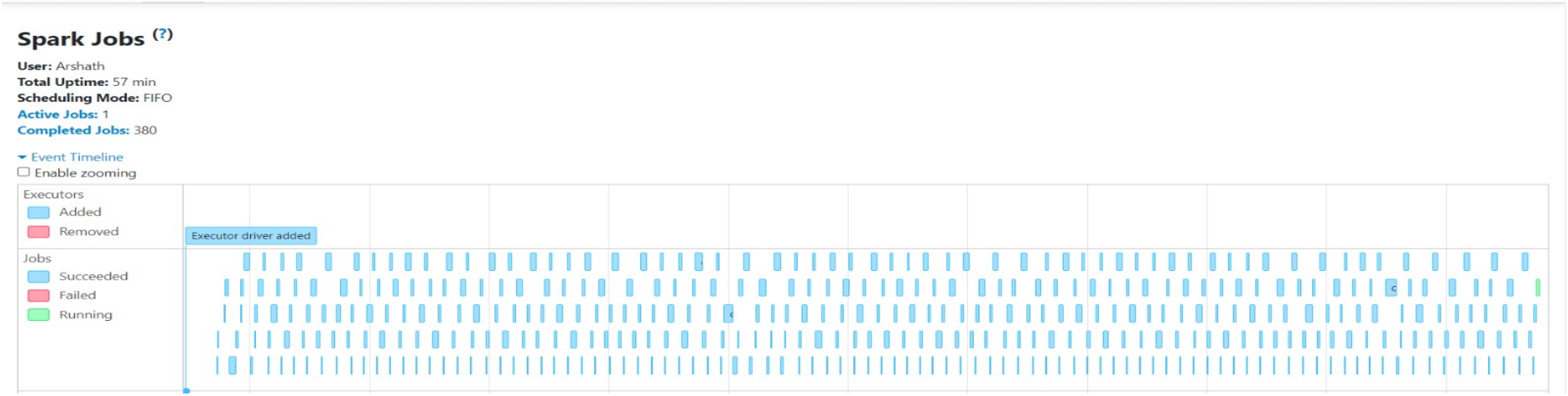

Figure: Spark UI

Results - Contd.

	category	date	title	count	Score	MaxScore	neg_score	neu_score	pos_score	wt_neg	wt_neu	wt_pos
0	EVs	2022-08-01	@1stMarsColonist @Tesla Ive seen phantom cars ...	1	[[0.1481754 0.7818548 0.06996979]]	0	0.148175	0.781855	0.06997	0.148175	0.781855	0.06997
1	EVs	2022-08-01	@TimOBrien @stilgherrian Signed up for the KIA...	1	[[0.19245706 0.6386235 0.16891949]]	0	0.192457	0.638624	0.168919	0.192457	0.638624	0.168919
2	EVs	2022-08-01	@ForwardBike @Howlrunner82 @ThunderWolf08 its ...	1	[[0.34761974 0.61530894 0.03707127]]	0	0.34762	0.615309	0.037071	0.34762	0.615309	0.037071
3	EVs	2022-08-01	"All that was great in the past was ridiculed,...	2	[[0.04469169 0.25750583 0.6978025]]	1	0.044692	0.257506	0.697802	0.089383	0.515012	1.395605
4	EVs	2022-08-01	OH: 10-year-boy to 9-year-old girl in Oakland ...	1	[[0.06985819 0.7173903 0.21275154]]	0	0.069858	0.71739	0.212752	0.069858	0.71739	0.212752

Figure: Aggregated Data with Sentimental Scores

```
[31] from scipy.stats import pearsonr
     list1 = time_df['t-1_pos_count']
     list2 = time_df['close_y']
     corr, _ = pearsonr(list1, list2)
     corr
```

```
0.17894162749047615
```

```
from scipy.stats import spearmanr
list1 = time_df['t-1_pos_count']
list2 = time_df['close_y']
corr, _ = spearmanr(list1, list2)
corr
```

```
0.16368557376227966
```

Figure: Correlation coefficient previous day positive sentiments count and close value of the current day - Tech Category

Results - Contd.

```
from scipy.stats import pearsonr
list1 = time_df['t-1_neg_count']
list2 = time_df['close_y']
corr, _ = pearsonr(list1, list2)
corr
```

```
-0.40180430845086584
```

```
from scipy.stats import spearmanr
list1 = time_df['t-1_neg_count']
list2 = time_df['close_y']
corr, _ = spearmanr(list1, list2)
corr
```

```
-0.470427713058414
```

Figure: Correlation coefficient prev day negative sentiments count and close value of the current day - EV Category

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