Stocks Price prediction with twitter sentiment analysis

BY
Awodaisi Adeyanju

August 2021

Outlines

- **Introduction**
  - Introduction
  - Research Aims and Objectives
  - Work done in this research
  - Justification for twitter preference over other social media
  - Related Modules
- **Review of related works**
- **Methodology**
  - Dataset collection and cleaning
  - Data Cleaning and preparation
  - Explanatory Analysis
  - Analysis of the pre-processed Datasets
- **Result Discussion**
- **Conclusion and Recommendation**
- **Legal/Ethic relating to the research.**
- **Project management Dashboard**

Introduction

- Prediction of stock market prices in the past using only historical price data have shown to be inadequate due to unsatisfactory accuracy level of the results. This shows that external factors other than historical data contribute to the prices' movement of stocks. Sentiments that are expressed daily by social medial users may contain some factors that could be analyzed and utilized to increase the accuracy of stock market price movement prediction.

Introduction - Research Aims and Objectives

Research Question

What is the effectiveness of Using Twitter Sentiment analysis for prediction of Stock price movement when compared with LSTM network algorithm?

Aim

To develop and suggest a more robust and reliable algorithm for stock market price movement predictions.

Introduction - Research Aims and Objectives

- **Objectives-** To analyze and compare the results of predictions of stock market prices movements using sentiment analysis and LSTM networks.

- To achieve the objectives the under listed steps were followed:
- I extracted raw sentiments from twitter about the HSBC and IAG stocks, performed sentiment analysis of the extracted tweets. We examined the correlation between the sentiments and the price movement of the stocks within the period covered.
- I extracted historic price data from yahoo finance for the selected stocks above within the same period, I used LSTM network to predict the next day price based on the extracted historic price data.
- And finally, I compared the results of the two steps above for model assessment and effectiveness.

Introduction - Work done in this research

- ✓ In this research, I used Snscrape library in python to collect tweets that contain sentiments about HSBC and IAG (using British airways tweets). The sentiments from the tweets extracted were processed further and used to predict the stock prices for the two listed companies.
- ✓ Furthermore, the effectiveness of this method was compared with the results of prediction using LSTM (Long short-term memory) model for the same prediction.
- ✓ The LSTM prediction model because it is presently considered as the most effective for stock price prediction using historical price data.

Introduction - Twitter preference over other social media

❖I opted for twitter over other social media because of the following reasons

- ✓Statista listed twitter among top 10 social media platform used by residents of United Kingdom.

- ✓Research mentioned that twitter cut across all ages, and it is ranked highest in terms of the percentage of educated followers among other social media [4].

Introduction - Related Modules

- This project afforded me the opportunity to put to test experience gained in the following modules

- ✓ Information retrieval
- ✓ modelling,
- ✓ And artificial neural networks modules.

Review of related works

- **Kalyani Joshi and Prof. Bharathi:** in 2017 extracted and analyzed news sentiments of financial articles from google, reuters and yahoo finance about apple stock. Though, they were able to achieve more than 50% accuracy in their prediction, my review of the research showed limitations in the following areas

- **Limitations**

✓ The sentiments analyzed were limited to financial related sentiments – The research made use of financial sentiment dictionary for its polarity classification.

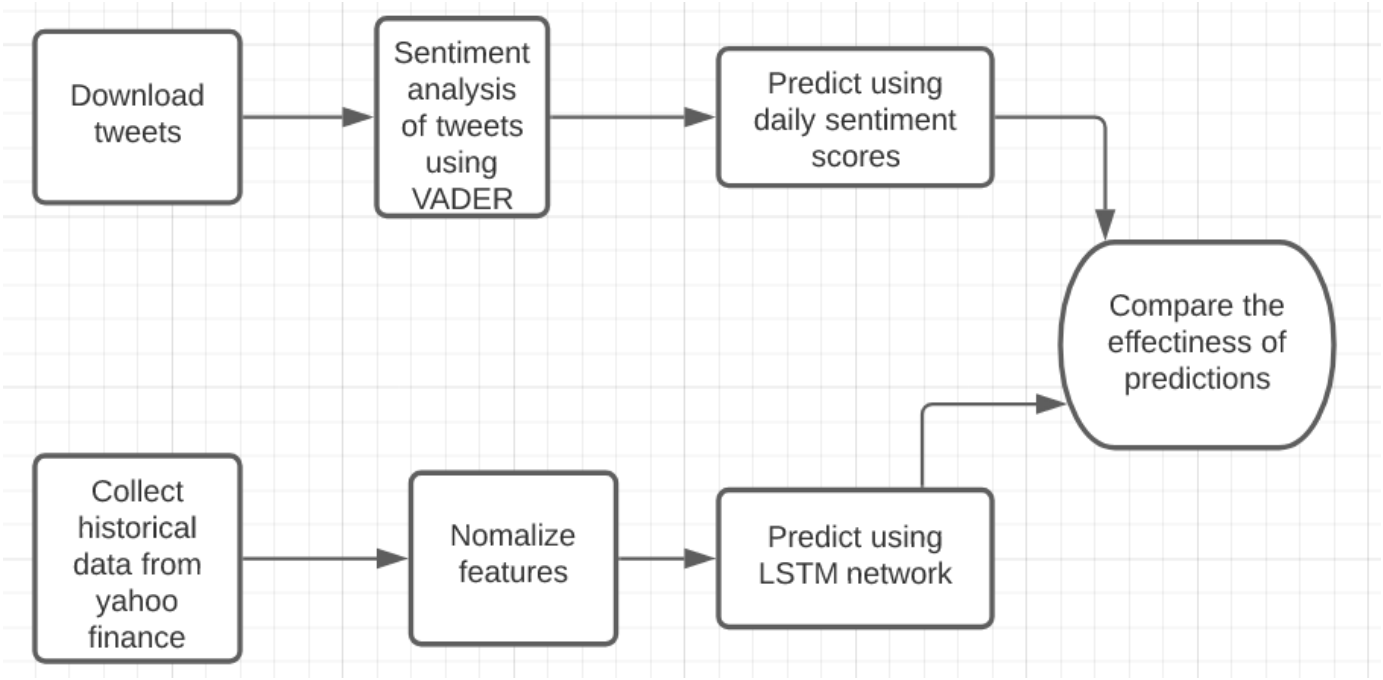
✓ Using only financial articles for sentiments analysis limits the sentiments to historical data and financial performance of the referenced stocks. Other external factors such as non-financial perception about the company may not be considered in the sentiment classification. – The comments around these articles are also embed in tweets sentiments

Review of related works

- **Sidra Mehtab & Jaydip Sen (April 2020) and Qiu J, Wang B, Zhou C (2020):** worked on the prediction of stock prices using historical prices data. Both made use of LSTM algorithms for the predictions. Sidra Mehtab & Jaydip Sen focused on every five minutes performance of stock price instead of usual dependance on daily close price for their prediction. Qiu J & co worked on the LSTM architecture for the prediction.

- **Limitation:**

✓ Prediction of stocks price with only historical price data may not provide accurate prediction as other factors beyond the financial data also affect movement of stocks.



Methodology - The summarizes Diagram for methodology Steps.

Methodology - Dataset collection and cleaning

❖ I made use of two datasets – tweets and historical prices data for the HSBC and IAG.

❖ **Tweets download**

✓ I made use of snscrape python library to extract all tweets in which HSBC and British Airways were mentioned between January and June 2021. BA was used for IAG because there were little tweets about IAG, anything sentiment about BA affects IAG directly. Also, there was another company named IAG that is into agriculture in the USA that may increase noise in our data. We downloaded 219,872 and 154,391 tweets respectively for HSBC and BA-IAG

❖ **The historical price data**

✓ I scraped from yahoo finance (fyahoo) with the help of free fyahoo API. Because HSBC and IAG are also listed in other markets (Hong Kong, Madrid, and New York), we used HSBC Ticker (HABS.L) for London stock exchange for the download.

Methodology - Dataset collected

	Datetime	Tweet Id	Text	Username	location	followers	likecount	retweetcount
0	2021-06-29 23:57:23+00:00	1410024629326860292	Is this a warning to get your money out of HSB...	deanprocter	Tributary to an ocean of Peace	2793	0	0
1	2021-06-29 23:57:22+00:00	1410024625761832961	@EzequielGN95 En línea: Banorte, Banamex y San...	AtletideSanLuis	Estadio Alfonso Lastras	158462	8	1
2	2021-06-29 23:54:57+00:00	1410024016329515008	@dgcoumans HolaDaniel, una disculpa por esta e...	HSBC_MX		220477	0	0
3	2021-06-29 23:52:52+00:00	1410023493161394180	@brayanceciliom2 Por favor compártenos por Men...	HSBC_MX		220477	0	0
4	2021-06-29 23:51:44+00:00	1410023206921113604	Huawei CFO says HSBC emails disprove basis for...	dev_discourse	National Capital Region	93361	0	0
...	...	...	...	...	...	...	...	...
219867	2021-01-01 00:11:19+00:00	1344798320590372864	hsbc, thx for the penny! umm i liked it moar ...	_bitcoiner		8074	7	1
219868	2021-01-01 00:03:21+00:00	1344796314534555648	@Mehdi_hsbc @Mediavenir Vrai	GrotaLaLegende		41	0	0
219869	2021-01-01	1344795673447854080	@Mediavenir Au moins ils vivront pas l'enfer d...	Mehdi_hsbc	Paris	311	3	0

1620abb3b148acb2ad2749e30ed61999e

	Datetime	Tweet Id	Text	Username	location	followers	likecount	retweetcount
0	2021-06-29 23:37:56+00:00	1410019736688148484	Bruselas abre una investigación sobre la compr...	PrimariosES	Madrid	9605	0	0
1	2021-06-29 23:21:50+00:00	1410015681932898312	Una chica me habló para hacer el examen prácti...	yrapte_	Banfield	900	1	0
2	2021-06-29 23:10:40+00:00	1410012871916933122	2021-06-29 às 20 horas : Temperatura: 11,5°C; ...	estacao_IAG	Água Funda - São Paulo - SP	2137	0	0
3	2021-06-29 23:10:01+00:00	1410012709215784961	2021-06-29 às 19 horas : Temperatura: 12,0°C; ...	estacao_IAG	Água Funda - São Paulo - SP	2137	0	0
4	2021-06-29 22:54:23+00:00	1410008777059938307	Te invitamos este jueves 01/07/2021 al webinar...	IAG_Guatemala	Guatemala City, Guatemala	199	2	1
5	2021-06-29 22:21:59+00:00	1410000622804475909	$IAG BUY\n\nIAMGOLD: Valuation Starting To Get...	Maggers78	Texas, USA	1548	3	0
6	2021-06-29 22:18:36+00:00	1409999772157726727	@IAG_Cargo @Kuehne_Nagel @NesteGlobal @SSieppi...	BiofuelsCent	Netherlands	27	0	0
7	2021-06-29 22:18:28+00:00	1409999736132804614	IAG Cargo Partners With Kuehne+Nagel and Neste...	BiofuelsCent	Netherlands	27	2	0

d89556d7932ac84c582b5c100588a48d11

Methodology - Dataset collected

	Date	Open	High	Low	Close	Adj Close	Volume
0	2021-01-04	165.949997	165.949997	149.550003	149.850006	149.850006	61675292
1	2021-01-05	147.000000	153.100006	144.250000	149.449997	149.449997	41207339
2	2021-01-06	151.399994	159.600006	148.600006	158.100006	158.100006	56332554
3	2021-01-07	159.550003	159.949997	151.750000	157.350006	157.350006	65284264
4	2021-01-08	159.149994	161.899994	155.636993	156.750000	156.750000	50877501

```
#IAG Price Historical Data
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2016-01-04	602.0	607.046997	593.500000	596.0	432.378662	7615055
1	2016-01-05	599.5	615.703979	598.969971	611.0	443.260681	6126552
2	2016-01-06	614.0	614.145020	599.200012	606.5	439.996063	7672847
3	2016-01-07	593.5	598.500000	577.856018	590.0	428.025909	8536283
4	2016-01-08	595.0	604.323975	586.116028	590.5	428.388672	4880253

```
#HSBC Price Historical Data
```

Methodology - Data Cleaning and preparation

❖ To prepare the dataset for the analysis and predictions, the following were done on the dataset.

✓ Determining and Removal of non-influential tweets- filtered out tweets from UserID with less than 1000 followers and those tweets that did not have at least 1 like.

✓ Tokenisation of tweets

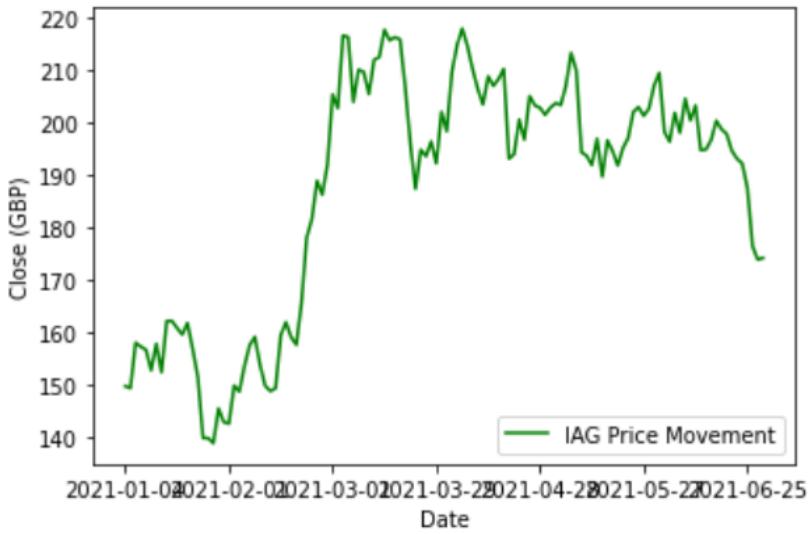
✓ Removal of non-english tweets

✓ Twitter Symbols Removal (“#”, “@” or htts)

✓ Stopwords

❖ We used the natural Language ToolKit library in python (NLTL) for the above exercises. Except for the removal of non-influential tweets. The extracted dataset for historical price data are cleaned.

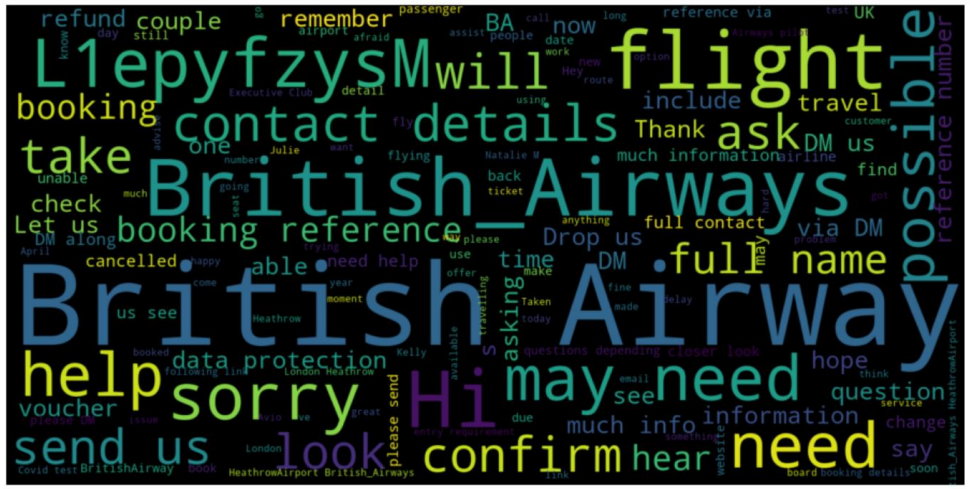
Methodology - Explanatory Analysis



- ✓ **Time series plots for the historical data:** The plots of the price's movements for HSBC and IAG shows similar properties
- ✓ The plots of the prices movements for HSBC and IAG shows similar properties, The prices of both stocks picked late January and went down again in February. The two stock reached their peaks in March 2021 before they went down again in June.

Methodology - Explanatory Analysis

- **Wordclouds plot of the tweets:** Some of the words in the tweets text that contributed to the polarity cores used for the prediction are shown below. For example, amazing, help, supporting, awards, great, and available were positive words that came through for HSBC. Negative words such as Extinction, rebellion, laundering could be seen as well. See figure below .



Methodology - Analysis of the pre-processed Datasets

❖ **Sentiment analysis of the tweet's dataset:** To get the polarity for each day between 1 January and 30 June 2021, I did sentiment analysis of each of the tweet's data for the stocks. I used VADER for the sentiment analysis of the tweets. VADER is a pretrained model in python that can classify or label social media lexicons into different polarity based on their semantic orientations. It allots polarity scores to each tweet based on their polarity depth.

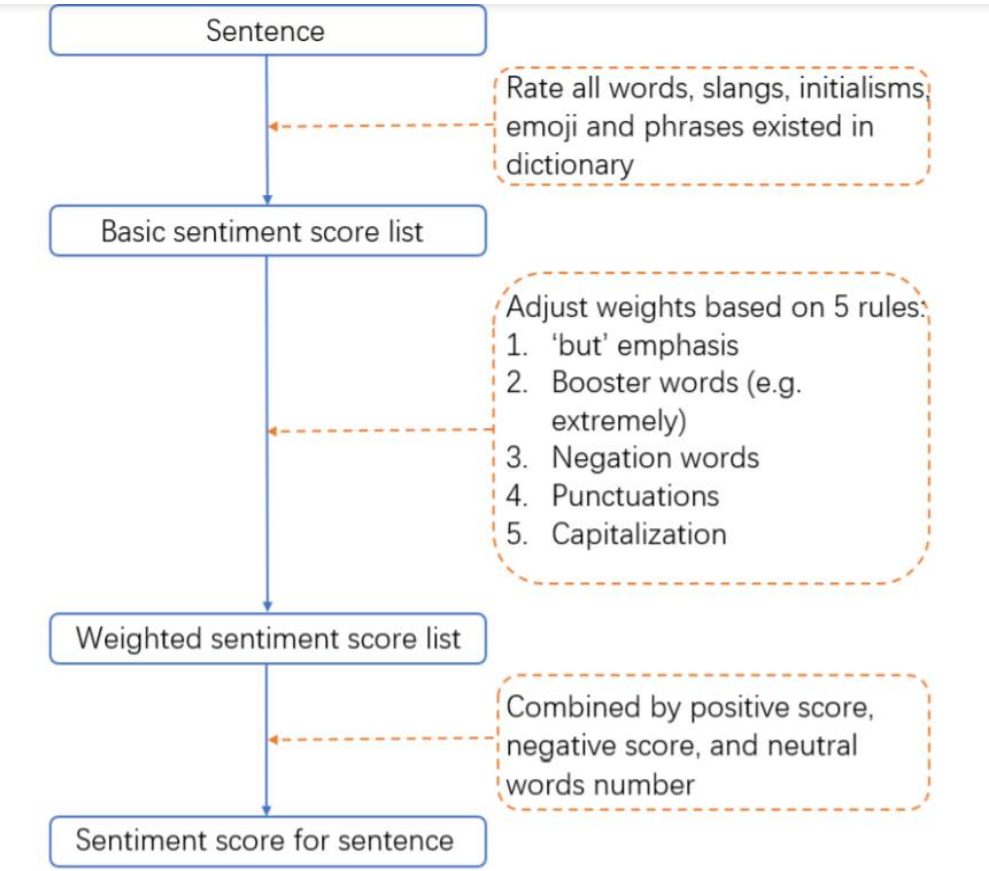
✓ Positive sentiment: (compound score >= 0.05)
Neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
Negative sentiment: (compound score <= -0.05).

✓ The score is allocated to show how negative, positive, or neutral a text/tweet is.

✓ Finally, VADER will calculate the average (compound) of the polarity score for each of the tweets.

Methodology - Analysis of the pre-processed Datasets

VADER sentiment package process Steps



Formular for the sentiment scores allocation for each sentence according to Heng Gui, (2019)

$$\text{Positive score} = \frac{\text{The sum of scores of all positive words}}{\text{sum of scores of all words + word count}} \quad \text{..... (Equation 1.3)}$$

$$\text{Negative score} = \frac{\text{The sum of scores of all Negative words}}{\text{sum of scores of all words + word count}} \quad \text{..... (Equation 1.4)}$$

$$\text{Neutral score} = \frac{\text{The count of all negative words}}{\text{sum of scores of all words + word count}} \quad \text{..... (Equation 1.5)}$$

$$\text{Compound score} = \frac{\text{The score sum of all words}}{\sqrt{(\text{The sum of scores of all words})\ 2+15}} \quad \text{..... (Equation 1.6)}$$

Methodology - Computation of compound score:

❖ **Datetime conversion and daily Computation of compound score:**

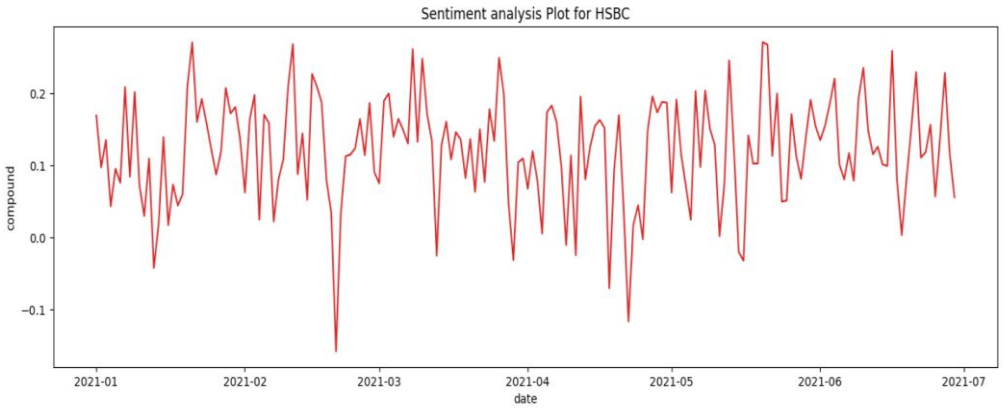
To have a similar dataframe that is similar to historical price data, we computed the daily average of the tweet sentiments compound polarity by using groupby function on python. See the result

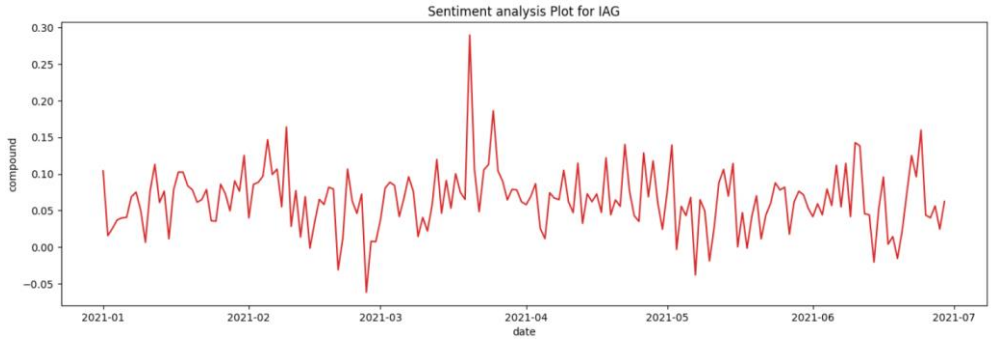
	Date	Close
0	2021-01-04	149.850006
1	2021-01-05	149.449997
2	2021-01-06	158.100006
3	2021-01-07	157.350006
4	2021-01-08	156.750000

date	description_lengths	compound
2021-01-01	30.266667	0.168983
2021-01-02	31.724138	0.097229
2021-01-03	32.714286	0.135231
2021-01-04	33.911111	0.043129
2021-01-05	32.510638	0.095304

Methodology - Analysis of the pre-processed Datasets

- Sentiment polarity plot for HSBC and IAG between January and June 2021.





Methodology - Analysis of the pre-processed Datasets

❖**Training of LSTM Model:**

✓**T**he historical price data was split into training (80%) and testing (20%) for each of HSBC and IAG. I normalized the training data to make them trainable in LSTM. I defined the datasets shapes of the dataset; the number of units and the dropout rate were defined. The dropout rate is defined to prevent overfitting, the output layers is specified to have a linear activation function. In addition, For the purpose of training, we used historical dataset between 1 January 2017 and June 2021.

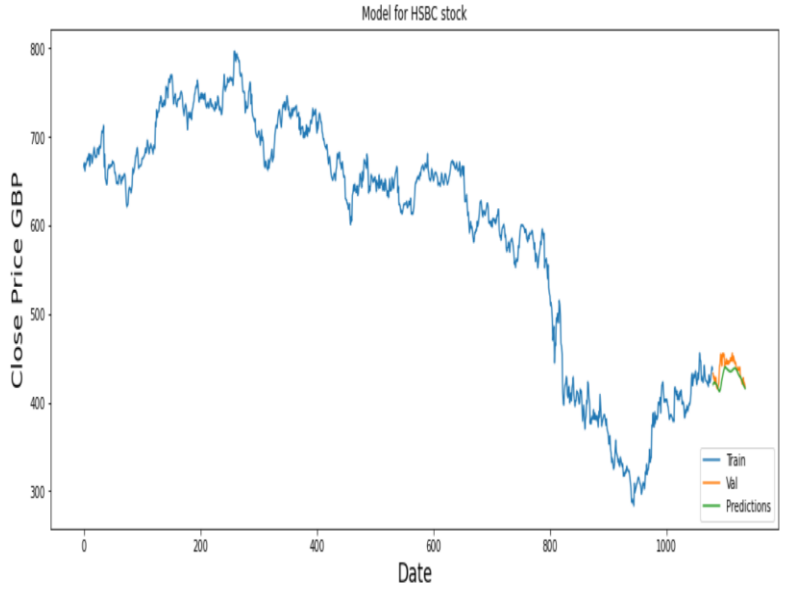
✓After the training of the models. I used both twitter Sentiment analysis and LSTM models to predict the price movements for the stocks of the two companies.

Result Discussion

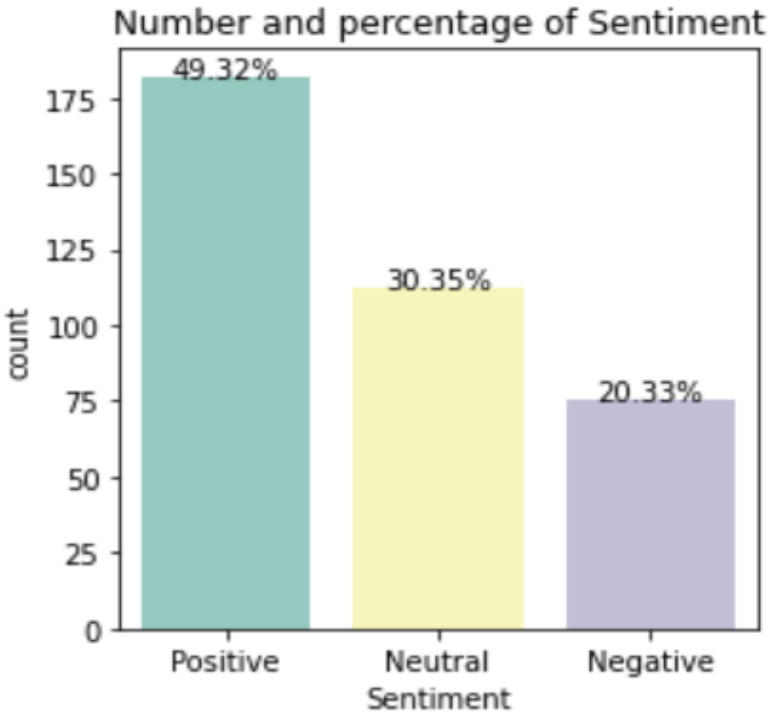
- ✓ When both LSTM network and Sentiment analysis were use for the predictions of the next day (1 July 2021 - 1135th day) stock price movements for both IAG and HSBC. The results showed that the predictions were correct using outcomes of sentiment analysis. The sentiment analysis predictions indicated positive prices movements for the stocks. However, prediction using LSTM model showed a positive price movement for HSBC and negative for IAG.

- ✓ It is also worthy of note that, although LSTM model prediction for IAG was negative as against the actual price which was positive , the prediction accuracy in terms of closeness to the actual price was 92.5%.

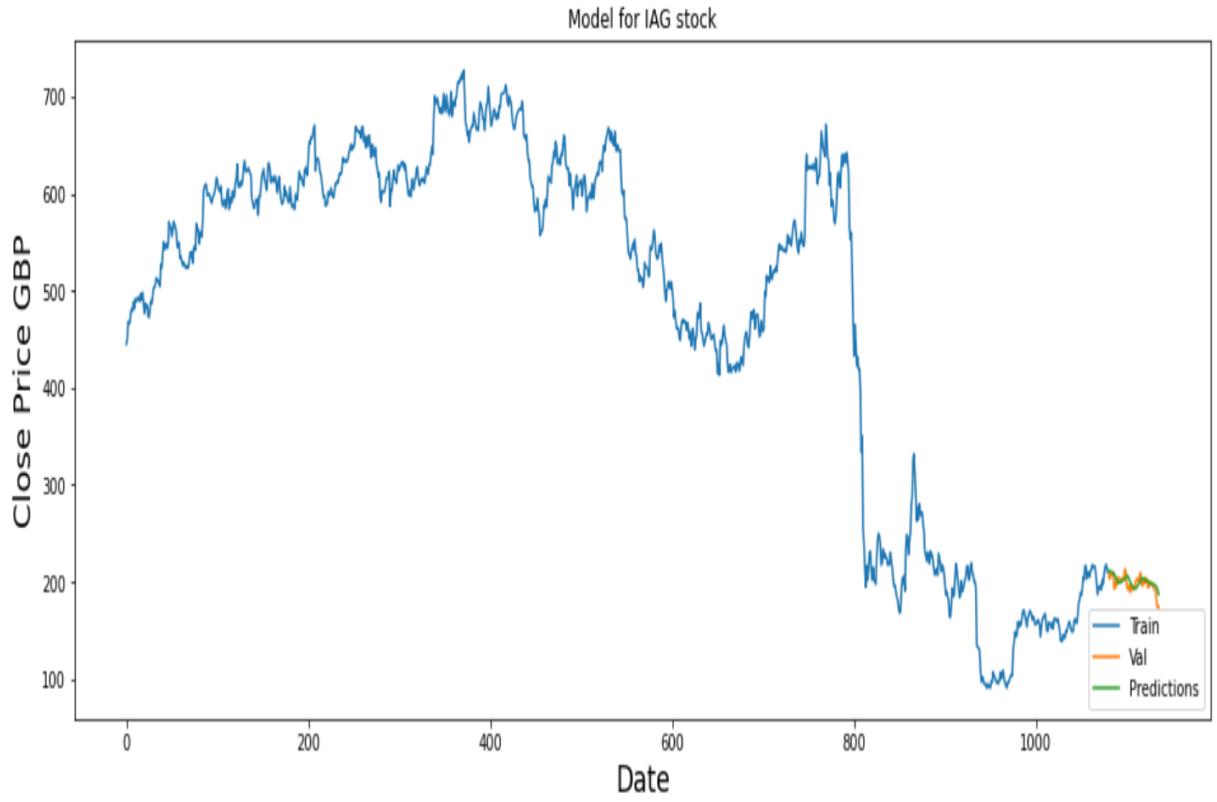
Result Discussion



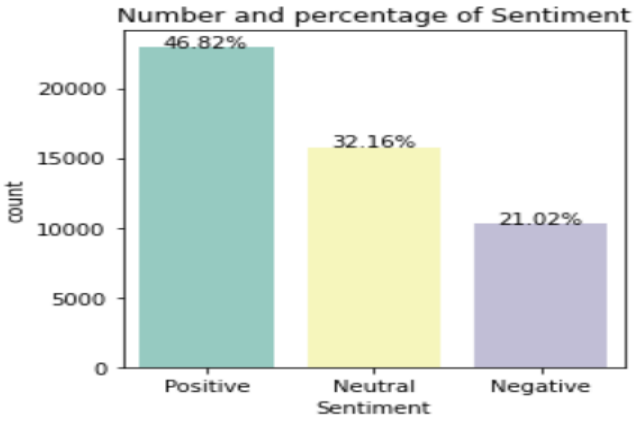
1133	176.399994	194.212250
1134	173.899994	191.263290
1135	174.220001	187.508408



Result Discussion – Continued
IAG Prediction results



1133	420.399994	418.468628
1134	421.299988	417.128601
1135	417.299988	415.940186



Conclusion and Suggestion

- ✓ I deduced from the results of the research that both methods could be employment for the predictions of stock prices movement.
- ✓ If properly analyzed, outcomes of sentiment analysis for the prediction may be good for classification prediction of stock price. The investors could know if stock price will move up or down.
- ✓ LSTM in another hand, will help for prediction of approximate amount in which a stock should be purchased.

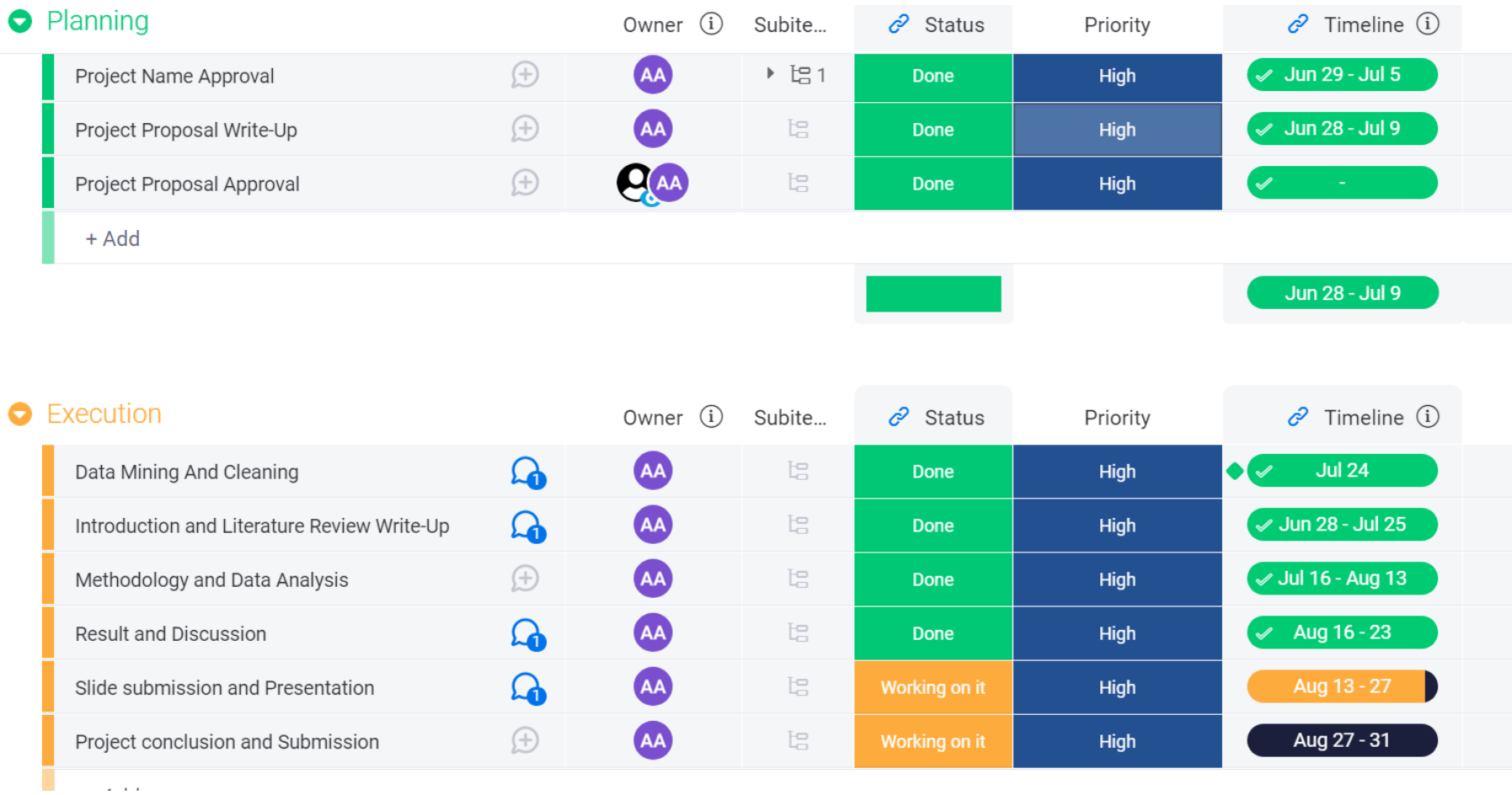
❖ **Further Work**

- ✓ Further work on combining the two methods in a system for price prediction will be a good research for scholars.

Legal/Ethic relating to the research.

- ✓ There was no legal issue related to the project.
- ✓ To avoid ethic violation in the project, I removed the Usernames and IDs of the twitter users. I did this to avoid the linking of any of the tweets to any individual or group.

Project Management Dashboard



Planning	Owner	Subite...	Status	Priority	Timeline
Project Name Approval	AA	1	Done	High	Jun 29 - Jul 5
Project Proposal Write-Up	AA		Done	High	Jun 28 - Jul 9
Project Proposal Approval	AA		Done	High	-
+ Add					
					Jun 28 - Jul 9

Execution	Owner	Subite...	Status	Priority	Timeline
Data Mining And Cleaning	AA		Done	High	Jul 24
Introduction and Literature Review Write-Up	AA		Done	High	Jun 28 - Jul 25
Methodology and Data Analysis	AA		Done	High	Jul 16 - Aug 13
Result and Discussion	AA		Done	High	Aug 16 - 23
Slide submission and Presentation	AA		Working on it	High	Aug 13 - 27
Project conclusion and Submission	AA		Working on it	High	Aug 27 - 31

Reference

- [1] Anshul Mittal and Arpit Goel (2014). Stock Prediction Using Twitter Sentiment Analysis http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf. Accessed 31 July 2021

- [2] Kalyani Joshi1 and Prof. Bharathi, (2017). Stock Trend Prediction using News sentiment analysis. A paper submitted to European Journal of Molecular & Clinical Medicine (5060 – 5069),Volume 07, Issue 02, 2020.

- [3] Michael Phi, (September 2018). Illustrated Guide to LSTM's and GRU's: A step by step explanation. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21. Accessed 30 June 2021.

- [4] statistical, (Oct 2020). Internet usage worldwide. https://www.statista.com/topics/1145/internet-usage-worldwide/#:~:text=In%202019%2C%20the%20number%20of,currently%20connected%20to%20the%20internet. Accessed 29 June 2021.

THANK YOU