Github Link:

https://github.com/anjanshrestha123/Sentiment-Analysis-for-Stock-Price-Prediction

Project Title: **Sentiment Analysis for Stock Price Prediction**

Team Members: **Anjan Shrestha, Aashish Pandey, Umesh Jaiswal, Puja Dhungana**

Project Description:

1. **Motivation**

Due to the availability of large data sets, Natural Language Processing (NLP) techniques and Machine Learning (ML) models have become an integral part of finance research. Many researchers have analyzed the text documents to see how investors' sentiment affects the stock price movement. Twitter data was used to predict the investors' mood and used it to analyze the stock markets movements (Mittal and Goel) [1]. They formed a naïve portfolio management strategy for their analysis and obtained 75.56% accuracy using Self Organizing Fuzzy Neural Networks on the Twitter feeds. Bollen et al. (2011) [2] predicted the stock market with 87% accuracy using a similar technique. Similarly, Kim et al. (2018) [3] used textual data from blogs, financial reports, and news to predict the stock price movements. Specifically, they used 8-K financial reports of firm's sector by sector and found that their approach improved the stock price prediction by 25.4%.

Although many managers and academic researchers have published their works on stock prediction, it is still identified as an important empirical problem in the finance field. In this paper, we will use data from different social media to predict public mood and stock price movements. We will also compare our study with some previous works to see if our approach improves prediction performance.

2. **Significance**

Stock market follows a random pattern and contains many calculated and uncalculated risks for the investors. This makes the stock market more fragile. The stock market sentiment is directly impacted by different factors such as politics, news and industry. Social media has become a common platform to share, discuss and give opinion on these factors. Thus, the positive and negative reviews on social media largely affects the stock market. Positive reviews increase the stock value while negative reviews decrease the stock value.

In this project we are analyzing such sentiment of people over social media to minimize some uncalculated risk. This could predict the possible future fluctuation of a stock price. It could attract many new investors. Moreover, it could help active investors to make decisions to enter or exit from the stock.

The stock market cannot be solely predicted by sentiment analysis. However, combining it with other fundamental analysis and technical analysis could help to make a more precise decision on particular stock.

3. **Objectives**

The main goal of this project is to predict stock prices based on market sentiment and provide a summary of people's sentiment using text summarization techniques. As we know, stock movement is largely based on people's feelings and emotions. So, if we can identify the correct emotions and sentiments in the market, there is a huge chance that we can predict the price of the stock market.

When people talk about stocks in blogs, chats and articles, they generally include positive sentiments such as happiness, hopefulness, enthusiasm etc. or negative sentiments such as disappointment, hate, sadness, etc. in their sentences. In the present world, social media has become the perfect platform to get those positive, negative or neutral sentiments from the people around the world.

So, our main objective is to extract those sentiments from different social media by using their APIs or by performing various web scraping techniques, to train and test our model out of those data, to predict the price of various stocks and to summarize findings in the form of text. In other words, our model should be able to suggest the best stocks that have a higher chance of price increase based on sentiment analysis.

4. **Features**

In this project, we will create a ML model to analyze the text data and perform the prediction on upward / downward movement of overall stock price (also the stock price for a list of highly volatile companies). We will exploit the sentiments expressed during communication or discussion in social media groups to perform the training and make predictions. Our data source will be various social networks like twitter, discord and facebook groups. We will also use the data from the New York Stock Exchange, publicly available in Kaggle [4] , to map the daily conversations with daily movement in stock price. Furthermore, we will generate the summary of conversations where the stocks are comparatively highly volatile by combining abstractive and extractive text summarization techniques.

Increment 1

Related Work (Background)

Stock market follows a random pattern and contains many calculated and uncalculated risks for the investors. This makes the stock market more fragile. With the advancement in data analysis, Natural Language Processing (NLP) techniques and Machine Learning (ML) models have become quite popular in stock market analysis. Previously, stock price was predicted by text summarization of available news articles, company's published financial and other reports. Currently, social media has become a common platform to share, discuss and give opinion on these factors. When people talk about stocks in blogs, chats and articles, they generally include positive sentiments such as happiness, hopefulness, enthusiasm etc. or negative sentiments such as disappointment, hate, sadness, etc. in their sentences. Thus, these positive, negative and neutral reviews on social media largely affect the stock price. Positive reviews increase the stock value while negative reviews decrease the stock value. The stock market sentiment is directly impacted by different factors such as politics, news and industry.

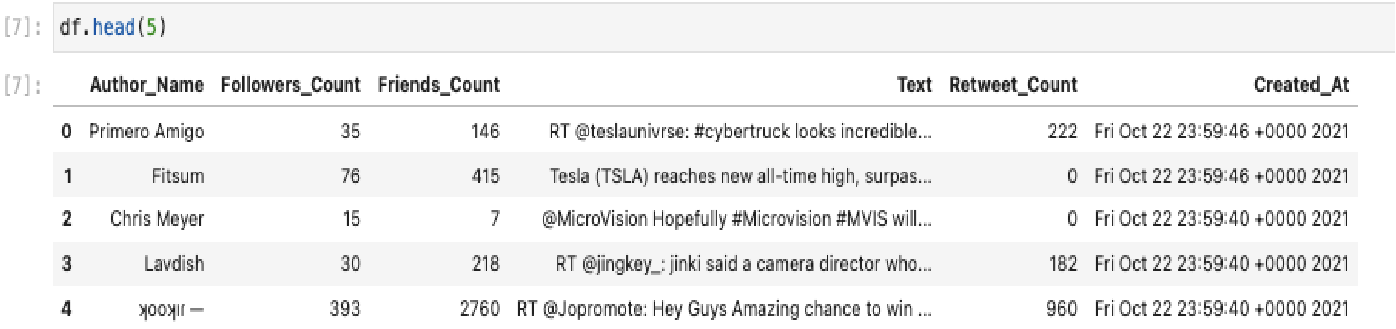
A study made by M. Naibpour and team compared nine machine learning models (Decision Tree, Random Forest, Adaptive Boosting (Adaboost), eXtreme Gradient Boosting (XGBoost), Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbor (KNN), Logistic Regression and Artificial Neural Network(ANN) and Recurrent Neural Network (RNN) and Long short-term memory (LSTM)) for four market groups (diversified financials, petroleum, non-metallic minerals and basic metal) from Tehran stock exchange showed RNN and LSTM model to perform best in their two methods. In the first method they used continuous data for features where Naïve-Bayes and Decision Tree showed least accuracy (approximately 68%) and RNN and LSTM showed high prediction (approximately 86%). In the next approach, they converted continuous data to binary data and used it in the ML model which increased accuracy to approximately 85% and 90% respectively (M. Nabipour, P. Nayyeri, H. Jabani, Shahab S., (Senior Member,IEEE), and Amir Mosavi. 1)[5]. Similarly, stock market prediction with linear regression done by K. Bhavsar predicted the stock price for next day [6]. Other research done on stock price using machine learning algorithms did not show high prediction values [7][8].

With time people's sentiments are being taken into consideration as part of stock price prediction research. Research in sentiment analysis on social media done by T. H. Nguyen and team obtained 54.41% average accuracy [9]. A prediction model based on Twitter sentiment analysis of Indonesian stock of 13 companies gave accuracy of 67.37% and 66.34% using random forest algorithm and naïve Bayes algorithm respectively for upcoming price fluctuation i.e., rise or fall [10]. Similarly, a distributive method for stock price prediction through sentiment analysis done by. M. Kim. and team showed improvement in prediction performance by 25.4% than baseline model [11].

As we can see, the predictions made by different models are either low or do not consider social media sentiment on stock price. The main objective of this project is to predict the fluctuation for the next day's price of a stock by analyzing the sentiments abstracted from Twitter using the Flair model. We are initially using Twitter dataset but we can use dataset from different other social networking sites like Facebook, Discord, Reddit, etc. This could increase the reliability of our system.

Dataset

For this project we used the dataset generated from Twitter using Twitter developer. Initially, we have worked with nearly 700 texts from Twitter API. We will increase the data size once the data is ready. We will also try to include the dataset from Facebook and Discord API; so that we could get more reliable and optimized output.



```
[7]: df.head(5)
```

[7]:

	Author_Name	Followers_Count	Friends_Count	Text	Retweet_Count	Created_At
0	Primero Amigo	35	146	RT @teslaunivrse: #cybertruck looks incredible...	222	Fri Oct 22 23:59:46 +0000 2021
1	Fitsum	76	415	Tesla (TSLA) reaches new all-time high, surpas...	0	Fri Oct 22 23:59:46 +0000 2021
2	Chris Meyer	15	7	@MicroVision Hopefully #Microvision #MVIS will...	0	Fri Oct 22 23:59:40 +0000 2021
3	Lavdish	30	218	RT @jingkey_: jinki said a camera director who...	182	Fri Oct 22 23:59:40 +0000 2021
4	ʞooʞıɾ —	393	2760	RT @Jopromote: Hey Guys Amazing chance to win ...	960	Fri Oct 22 23:59:40 +0000 2021

Fig: Raw dataset from twitter response

Yahoo Finance API is an open source with a huge range of data on stock, easy to set up and simple. Thus, after collecting the dataset we used Yahoo Finance API for labelling stock price. Text sentiment is mapped with the stock price of that day.

Details design of Features

After getting responses from Twitter API, there were a lot of features and out of those, we have selected six specific features that have the highest impact on the stock price i.e. tweet, follower's count, friend's count, retweet count and created date of tweet. Out of the tweet feature, we have used a flair model to get sentiment and its probability feature. Flair model gave us the sentiment in text format i.e. POSITIVE and NEGATIVE. We have encoded these text sentiment into encoded features i.e. 1 for POSITIVE and -1 for NEGATIVE. Also, sentiment probability features inform about accuracy of sentiment. And out of the created date of tweet feature, we have created a stock price label by calling the yahoo finance module.

- **Analysis**:

Initially, we have extracted tweets from 7 days form (Oct 23 2021 to Oct 30 2021) through twitter API. We analyzed approximately 700 tweets(100 from each day) and looked at the sentiment distribution of tweets in the dataset. Out of 696 text data, 355 were associated with Negative sentiment(encoded to -1) and 331 texts were associated with Positive sentiment (encoded to 1). The bar graph below shows the distribution of sentiment in our dataset.

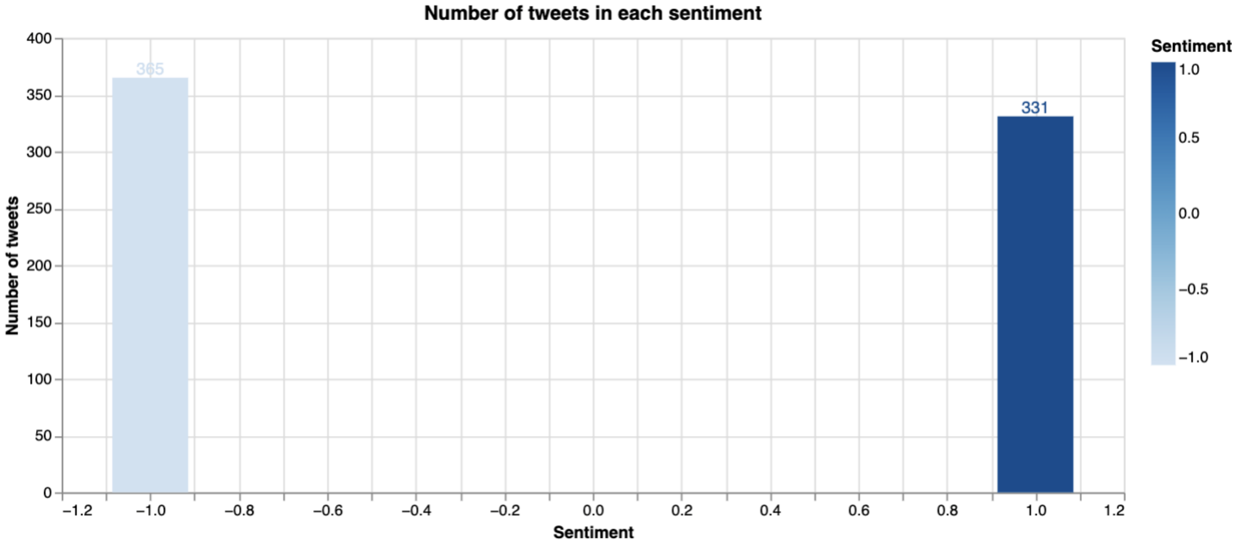


Fig: Sentiment Distribution in initial dataset.

Furthermore, we also analyzed the follower's count distribution of the account holders in the tweets. The power law distribution (long tail distribution) is absorbed in the diagram below. This shows that a very few of the account holders have a high number of followers and a lot of (most of the accounts) have a low number of followers.

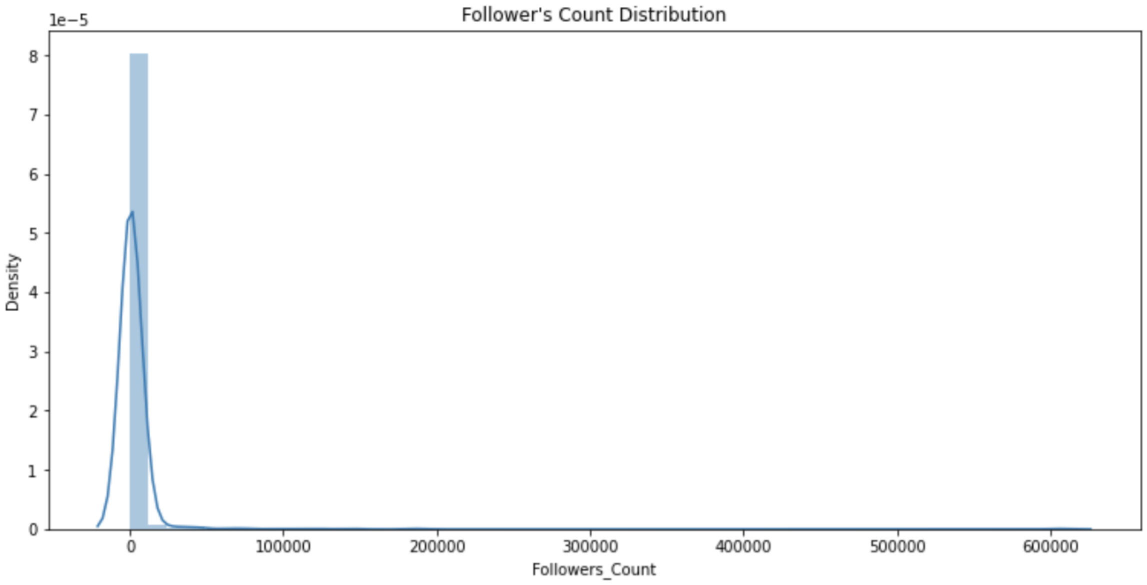
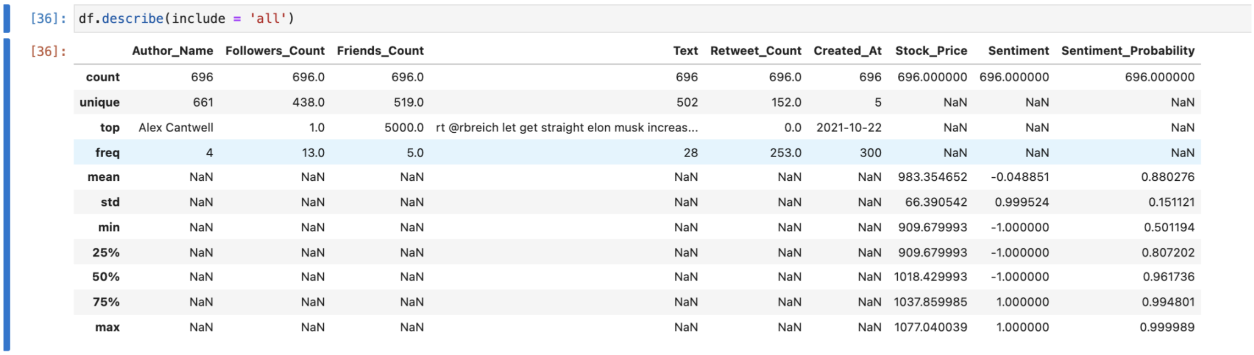


Fig: Followers count distribution in initial dataset

The table below shows the description of our dataset. We can visualize the different statistical distribution of the different attributes in our dataset below.



```
[36]: df.describe(include = 'all')
```

[36]:

	Author_Name	Followers_Count	Friends_Count	Text	Retweet_Count	Created_At	Stock_Price	Sentiment	Sentiment_Probability
count	696	696.0	696.0	696	696.0	696	696.000000	696.000000	696.000000
unique	661	438.0	519.0	502	152.0	5	NaN	NaN	NaN
top	Alex Cantwell	1.0	5000.0	rt @rbreich let get straight elon musk increas...	0.0	2021-10-22	NaN	NaN	NaN
freq	4	13.0	5.0	28	253.0	300	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	983.354652	-0.048851	0.880276
std	NaN	NaN	NaN	NaN	NaN	NaN	66.390542	0.999524	0.151121
min	NaN	NaN	NaN	NaN	NaN	NaN	909.679993	-1.000000	0.501194
25%	NaN	NaN	NaN	NaN	NaN	NaN	909.679993	-1.000000	0.807202
50%	NaN	NaN	NaN	NaN	NaN	NaN	1018.429993	-1.000000	0.961736
75%	NaN	NaN	NaN	NaN	NaN	NaN	1037.859985	1.000000	0.994801
max	NaN	NaN	NaN	NaN	NaN	NaN	1077.040039	1.000000	0.999989

Fig: Description of the dataset

Implementation

In this project, for stock market sentiment analysis we have extracted raw text data from Twitter API. Then we converted raw data to our desired dataset from API response i.e. from the pool of raw data of seven days we collected approximately 700 tweets of Tesla. After which, we used feature engineering for text cleaning, sentiment analysis, retweet counting and counting number of followers. For sentiment analysis we used a pre-trained Flair model. We used yahoo finance API for price labelling. Next, we created and trained a lstm and regression model to predict the stock price. Subsequently, we used text summarization to show our output.

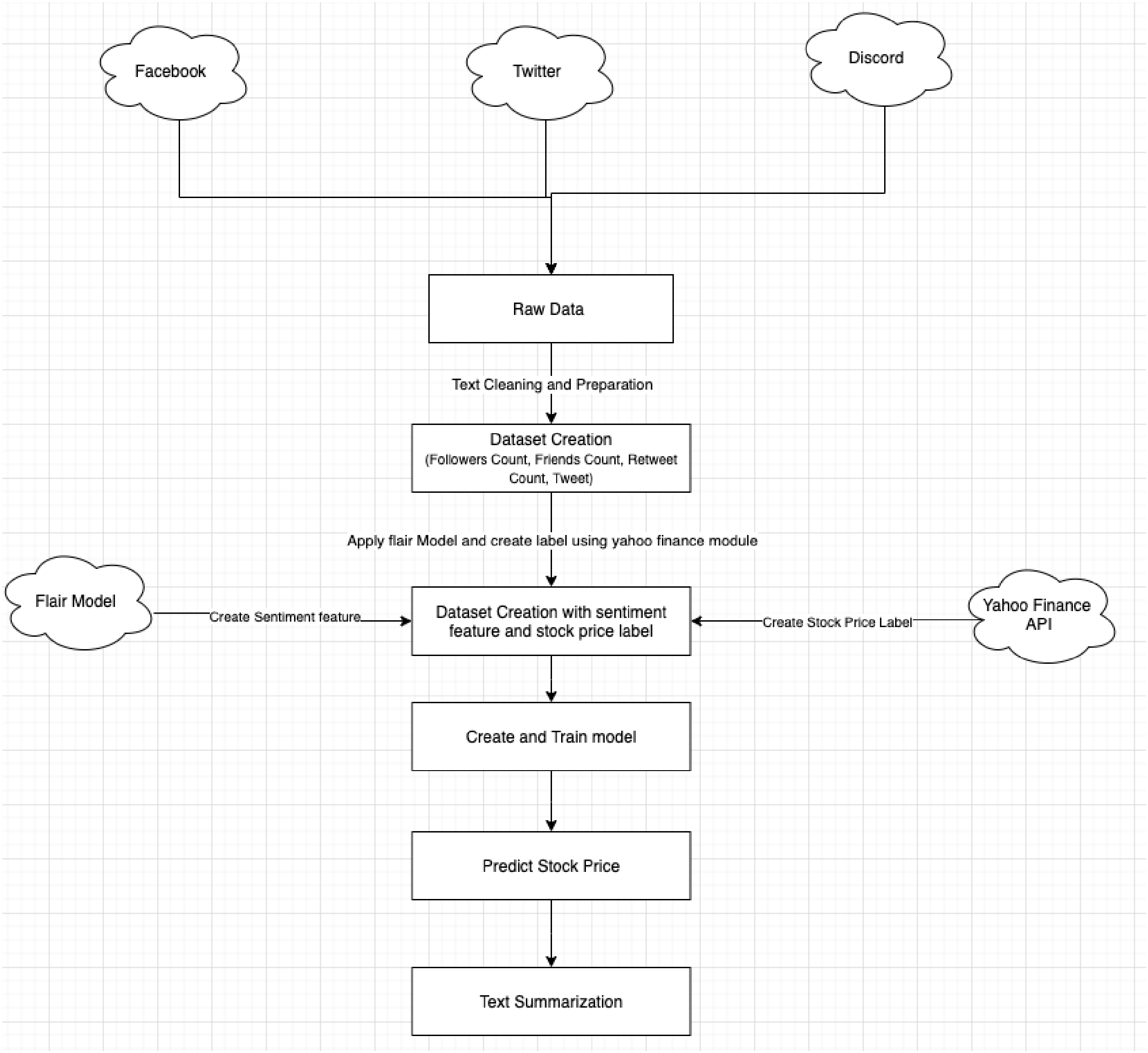
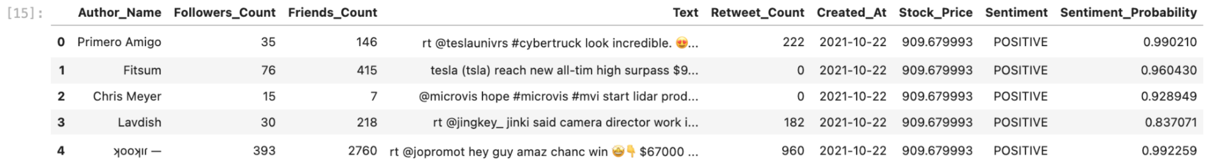

Facebook
Twitter
Discord
Raw Data
Text Cleaning and Preparation
Dataset Creation
(Followers Count, Friends Count, Retweet Count, Tweet)
Apply flair Model and create label using yahoo finance module
Flair Model
Create Sentiment feature
Dataset Creation with sentiment feature and stock price label
Create Stock Price Label
Yahoo Finance API
Create and Train model
Predict Stock Price
Text Summarization

Fig: Overall view of project design

- **Preliminary Results**:

We have successfully implemented the pre-trained flair model for the sentiment detection of the text. The flair model takes the text as input and returns the positive/negative label of the text with its degree(probability) value. We also have extracted the everyday stock price label from yahoo finance. We have also fetched various important features of each tweet like Followers Count of account holder, Friends count of account holder, retweet count etc from the twitter API. Finally, we have mapped the everyday texts and its features with everyday's stock price to create the final dataset for stock prediction.

The dataframe of these features is shown below:



[15]:

	Author_Name	Followers_Count	Friends_Count	Text	Retweet_Count	Created_At	Stock_Price	Sentiment	Sentiment_Probability
0	Primero Amigo	35	146	rt @teslaunivrs #cybertruck look incredible. 😍...	222	2021-10-22	909.679993	POSITIVE	0.990210
1	Fitsum	76	415	tesla (tsla) reach new all-tim high surpass $9...	0	2021-10-22	909.679993	POSITIVE	0.960430
2	Chris Meyer	15	7	@microvis hope #microvis #mvi start lidar prod...	0	2021-10-22	909.679993	POSITIVE	0.928949
3	Lavdish	30	218	rt @jingkey_ jinki said camera director work i...	182	2021-10-22	909.679993	POSITIVE	0.837071
4	ʞooʞıɾ —	393	2760	rt @jopromot hey guy amaz chanc win 🤩👇 $67000 ...	960	2021-10-22	909.679993	POSITIVE	0.992259

Also, we have started creating regression models. The models are yet to be fine tuned to generate the results. We expect to have a significant prediction result by the next increment.

• **Project Management**.

We created a WhatsApp group and discussed our progress every week. We also used Github to share our work. During the weekly meetings, we discussed the challenges that we could face during our project implementation. And, to keep the track of the workflow, we used trello as a kanban board.

o **Implementation status report**

Group members' Contribution

Group member	Contribution	Responsibility(completed)	Work to be done

Aashish Pandey	25%	dataset creation and feature engineering	train and evaluate the model
Anjan Shrestha	25%	raw data extraction	Create different regression model
Puja Dhungana	25%	creating sentiment features by using Flair model	apply abstractive summarization
Umesh Jaiswal	25%	related works and model design	generate the final prediction

Issues/Concerns:

We have yet to determine the time lag between the tweet date and change in stock price. For example, to determine if the tweet from Sunday affects the price of Sunday, or Monday or Tuesday. Once we have our final dataset and model ready, we will map the dataset with the correct label accordingly.

References/Bibliography:

1. Mittal,A., & Goel,A. Stock Prediction Using Twitter Sentiment Analysis
2. Bollen, J., & Mao, H. (2011). Twitter mood as a stock market predictor. Computer, 44(10), 0091-94

3. M. Kim, E. L. Park and S. Cho, "Stock price prediction through sentiment analysis of corporate disclosures using distributed representation", *Intelligent Data Analysis*, vol. 22, no. 6, pp. 1395-1413, Jan. 2018

4. Dominik Gawlik, New York Stock Exchange: S & P 500 companies historical price with fundamental data. *Kaggale,* version 3

5. M. Nabipour, P. Nayyeri, H. Jabani, Shahab S., (Senior Member,IEEE), and Amir Mosavi. Predicting Stock Market Trends Using. Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a comparative Analysis. *IEEEAccess. 2020*

6. K. Bhavsar. Stock Market Prediction with Linear Regression

7. H. Chung, K. Shin. Genetic Algorithm-Optimized Long Short-Term Memory Network for Stock Market Prediction

8. W. Long, Z. Lu, L. Cui. Deep Learning-Based Feature Engineering for Stock. Price movement Prediction

9. T. H. Nguyen, K. Shirai, J, Velcin. Sentiment Analysis on Social Media for Stock Movement Prediction. 2015

10. Trisedy BD. Stock Price Prediction Using Linear Regression Based on Sentiment analysis, *International Conference on Advanced Computer Science and Information System (ICACSIS). 2015*

11. M.Kim, E. L. Park, S. Cho. Stock Price Prediction Through Sentiment Analysis. of Corporate Disclosure using Distributed Representation. 2018