Sentiment Analysis and Abstractive Summarization For Stock Price Prediction

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

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

Presentation Video Link: https://drive.google.com/file/d/1SaU2UVIPrMO2pG-SIrZo1mt3FQqy4yDU/view?usp=sharing

Demo Video Link: https://drive.google.com/file/d/1ElgeFftKwYPkfwrYbA6jjOC9Q-6Lz0kM/view?usp=sharing

1. Project Description:

1.1 **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) (Mittal, 2012)[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) (*Twitter Mood Predicts the Stock Market*, 2021) [2] predicted the stock market with 87% accuracy using a similar technique. Similarly, Kim (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.

1.2 **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.

1.3 **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.

1.4 **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 from twitter. We will also use the data from the New York Stock Exchange, publicly available in Kaggle (*S&P 500 Stock Data*, n.d.)[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.

2. Increment 1

2.1 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 the next day (Bhavsar, n.d.)[6]. Other research done on stock price using machine learning algorithms did not show high prediction values (Chung & shik Shin, 2018)[7] (Long et al., 2018)[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 (Nguyen et al., 2015, #)[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 (Cakra & Trisedya, 2015)[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 [3 (Kim et al., 2018)].

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.

2.2 **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.

```
[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	ʞooʞıɹ —	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.

2.3 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.

2.4 **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.

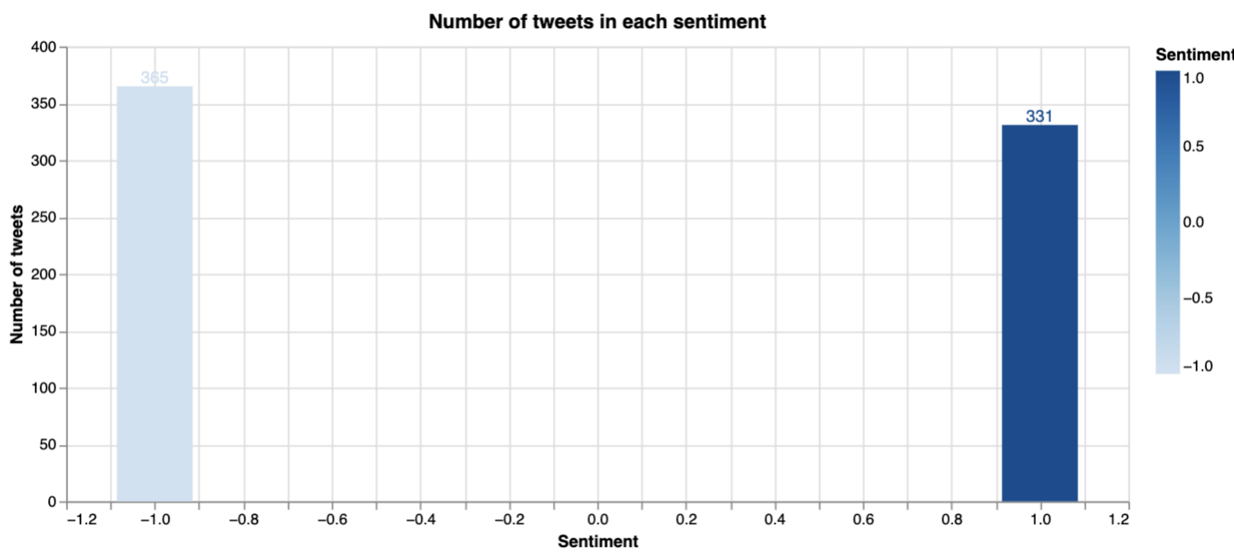

Fig: Sentiment Distribution in initial dataset.

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.

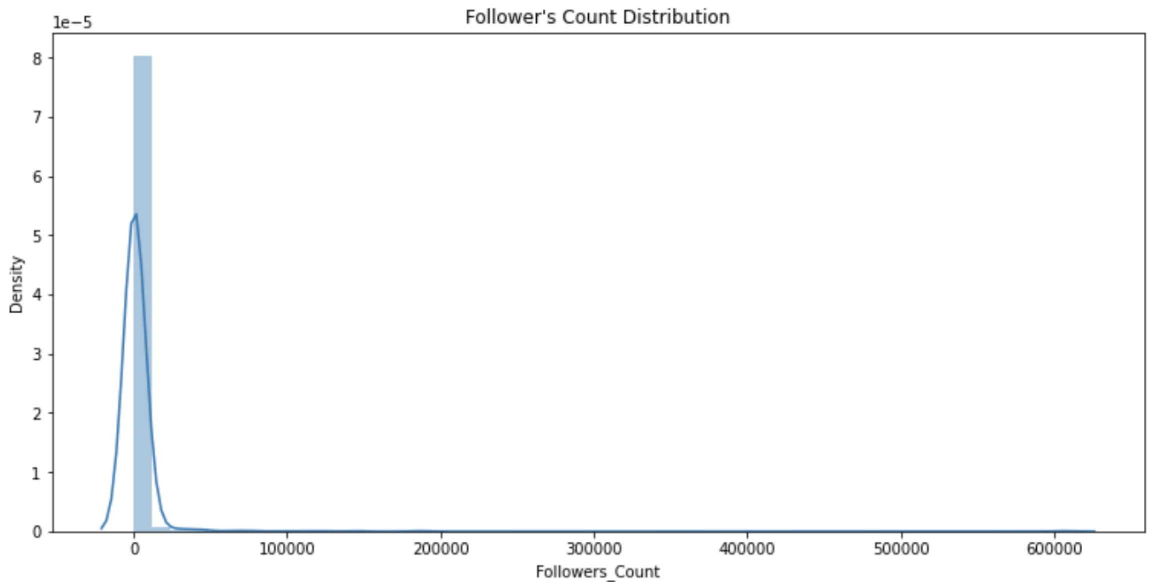

Fig: Followers count distribution in initial dataset

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

2.5 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.

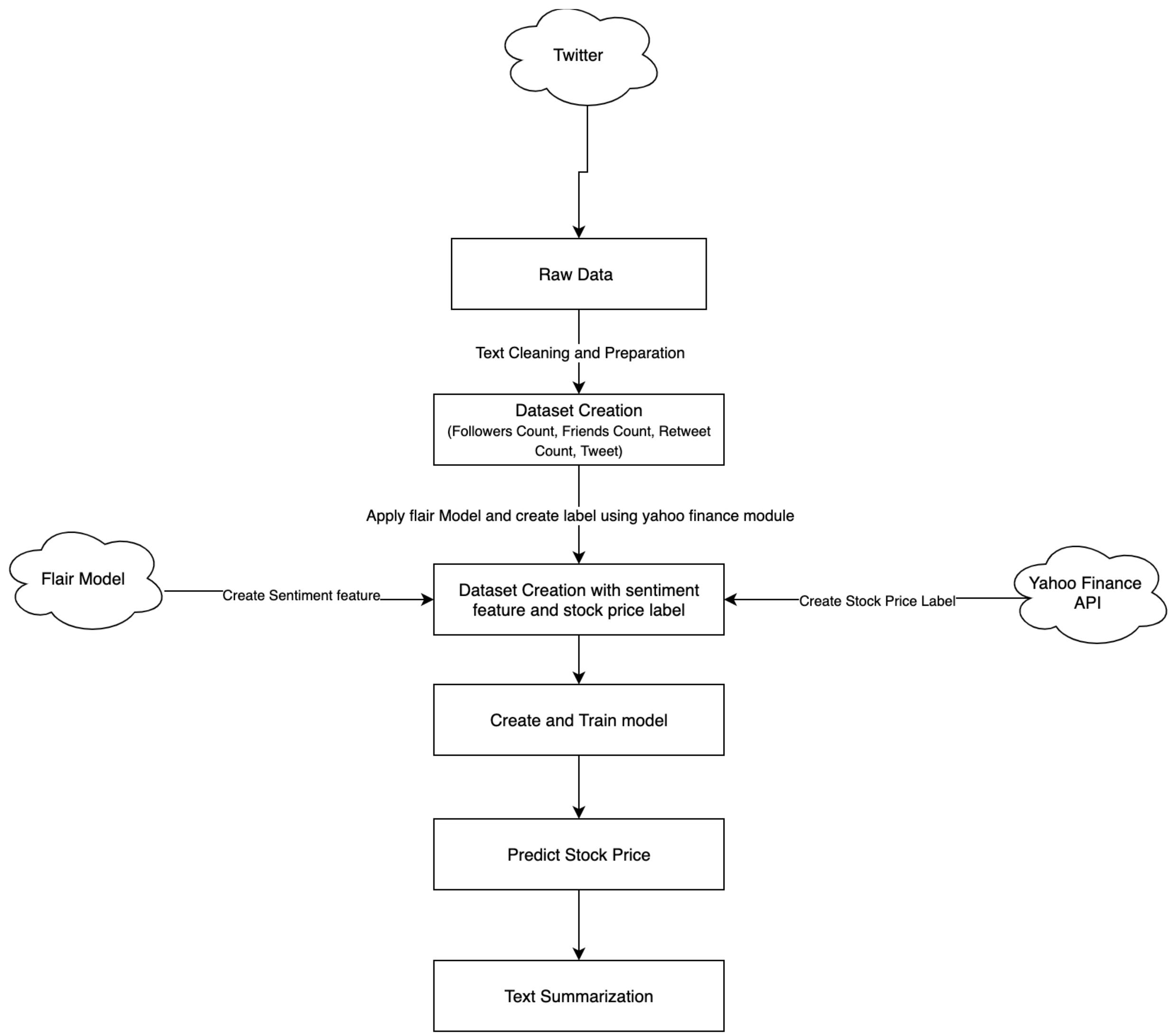

Fig: Overall view of project design

2.6 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	ʞooʞɹı —	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.

2.7 **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.

2.8 References/Bibliography:

1. A. Mittal and A. Goel, Stock Prediction Using Twitter Sentiment Analysis

2. J. Bollen and H. Mao, (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

3. Project - Final

3.1. Introduction

Stock market is an open platform and a good investment opportunity for everyone. However, people's participation in the stock trade is based upon how much risk a person can take. What if we could minimise these risks and make it a safe investment for general people.

The price of the stock fluctuates day by day. There are different factors that fuels this fluctuation like politics, technology changes, war or conflicts, natural calamities, people's sentiment, etc. These factors bring many calculated and uncalculated risks with them to the stock market. There are many research done nowadays for analysing these factors. Study of people's sentiment for the stock market prediction is one of the important topics of discussion. 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 project, 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.

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. 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. Positive reviews increase the stock value while negative reviews decrease the stock value. 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.

In this project, we analyze such sentiment of people over social media to minimize some uncalculated risk. Basically, we predict stock prices based on market sentiment and provide a summary of people's sentiment using text summarization techniques. The main goal of this project 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 predict the possible future fluctuation of a stock price based on sentiment analysis. It will attract many new investors. Moreover, it could help active investors to make decisions to enter or exit from the stock.

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 from twitter groups. We will also use the data from the New York Stock Exchange, publicly available in Kaggle [1] , 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.

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.

3.2 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 [2]. Similarly, stock market prediction with linear regression done by K. Bhavsar predicted the stock price for the next day [3]. Other research done on stock price using machine learning algorithms did not show high prediction values [4][5].

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 [6]. 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 [7]. 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 [8].

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.

3.3 Model

o Architecture Diagram with explanation

Predicting stock price using sentiment analysis is a supervised machine learning problem where regression is used. To make predictions, raw data is collected from Twitter API and a dataset is created out of it. The dataset created is further passed to the feature engineering process that uses yahoo finance API and flair module that creates features required to run and train our model. Using the features extracted, different machine learning models are created, trained, evaluated and the best regression model is used to make stock price prediction. Also, a pre-trained abstractive model is used to summarize the latest sentiment that is fetched from Twitter API.

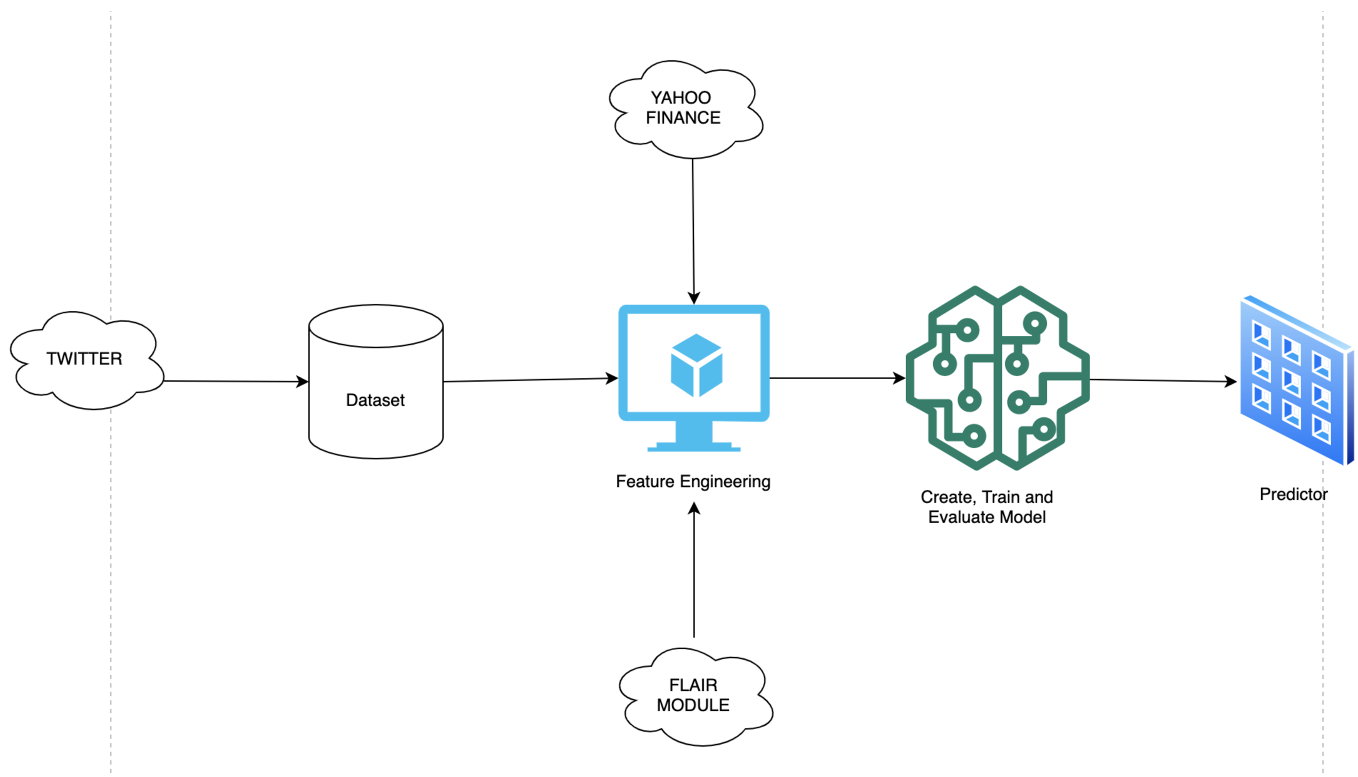

Fig: Architecture Diagram

o Workflow diagram with explanation

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.

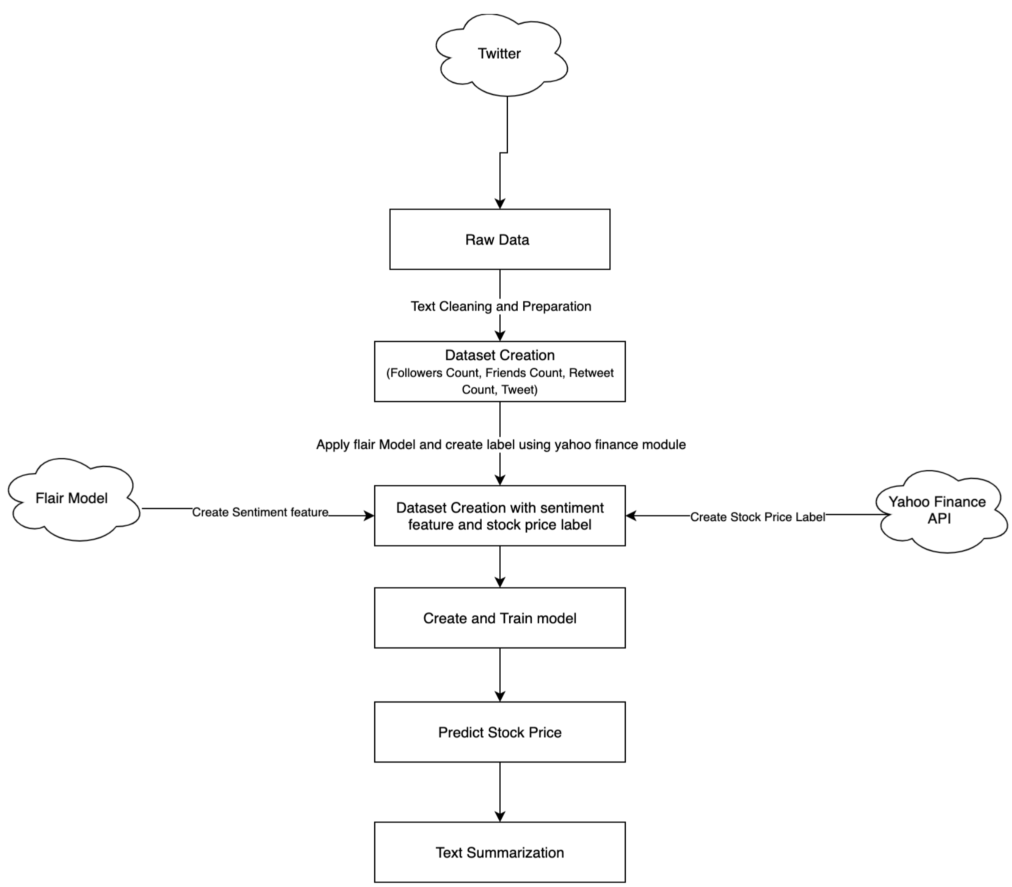

Fig: Overall view of project design

3.4 Dataset

o Detailed description of 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.

```
[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	ʞooʞıɹ —	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

o Detail design of Features with diagram

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.

After getting responses from Twitter API, there were around forty 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. The selection is made based on literature review.

```
▾ 99:
    created_at: "Fri Oct 22 23:55:51 +0000 2021"
    id: 1451698848758337500
    id_str: "1451698848758337537"
    full_text: "RT @1batron: @elonmusk #floki #FLOKI #FLOKIINU Floki loves Tesla Does Tesla love Floki? @elonmusk https://t.co/EBQgFofYPD"
    truncated: false
  ▸ display_text_range: [] 2 items
  ▾ entities:
    ▸ hashtags: [] 3 items
      symbols: [] 0 items
    ▸ user_mentions: [] 3 items
      urls: [] 0 items
    ▾ media: [] 1 item
      ▸ 0:
  ▾ extended_entities:
    ▸ media: [] 1 item
  ▸ metadata:
    source: "<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>"
    in_reply_to_status_id: null
    in_reply_to_status_id_str: null
    in_reply_to_user_id: null
    in_reply_to_user_id_str: null
    in_reply_to_screen_name: null
  ▸ user:
    geo: null
    coordinates: null
    place: null
    contributors: null
  ▸ retweeted_status:
    is_quote_status: false
    retweet_count: 7
    favorite_count: 0
    favorited: false
    retweeted: false
    possibly_sensitive: false
    lang: "en"
▾ search_metadata:
    completed_in: 0.108
    max_id: 1451699834151444500
    max_id_str: "1451699834151444481"
    next_results: "?max_id=1451698848758337536&q=tesla%20until%3A2021-10-23&lang=en&count=100&include_entities=1"
    query: "tesla+until%3A2021-10-23"
    refresh_url: "?since_id=1451699834151444481&q=tesla%20until%3A2021-10-23&lang=en&include_entities=1"
    count: 100
    since_id: 0
    since_id_str: "0"
```


Fig. Total Features available from Twitter API

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.

	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	1	0.990210
1	Fitsum	76	415	tesla (tsla) reach new all-tim high surpass $9...	0	2021-10-22	909.679993	1	0.960430
2	Chris Meyer	15	7	@microvis hope #microvis #mvi start lidar prod...	0	2021-10-22	909.679993	1	0.928949
3	Lavdish	30	218	rt @jingkey_ jinki said camera director work i...	182	2021-10-22	909.679993	1	0.837071
4	ʞooʞıɾ —	393	2760	rt @jopromot hey guy amaz chanc win 🤩👇$67000 ...	960	2021-10-22	909.679993	1	0.992259

Fig. Final Dataset after using the Yahoo Finance API and Flair Module

3.5 Analysis of data

o Data Pre-processing

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.

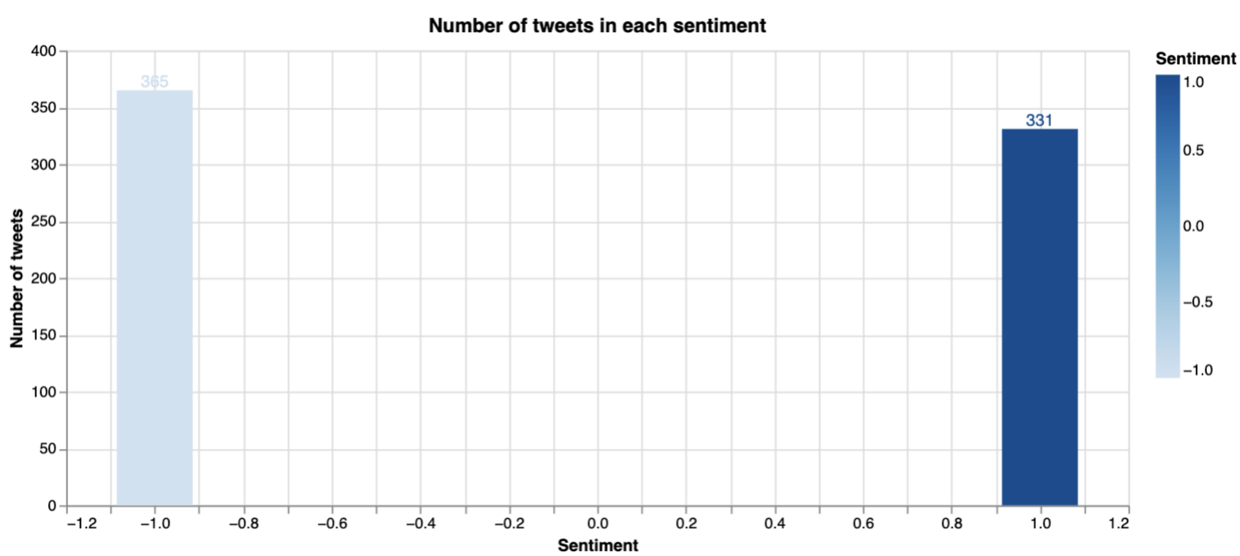

Fig: Sentiment Distribution in initial dataset.

o Graph model with explanation

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.

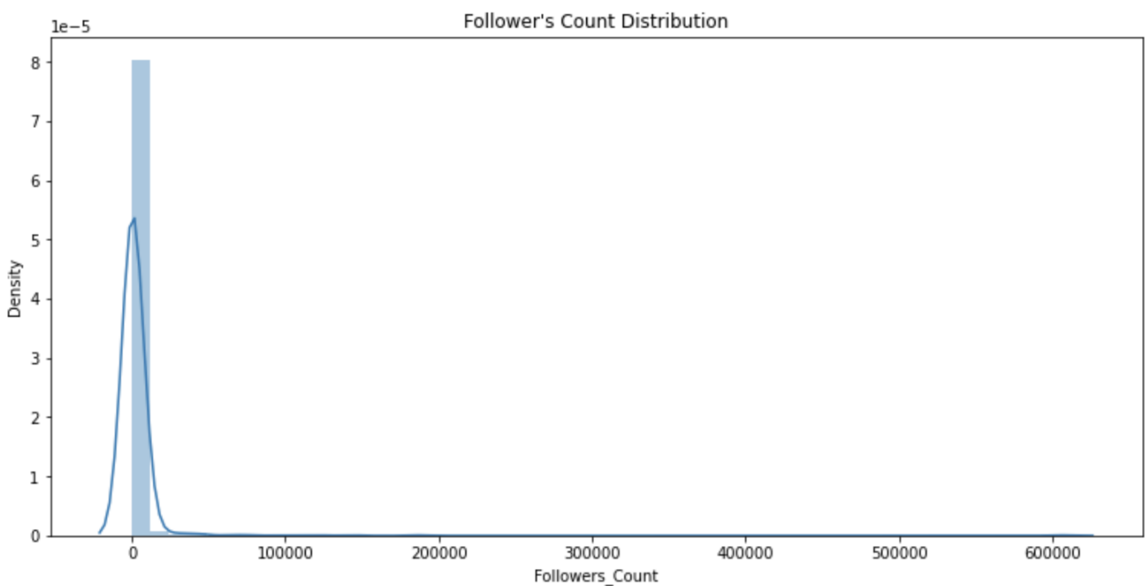

Fig: Followers count distribution in initial dataset

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

3.6 Implementation

o Algorithms / Pseudocode

The algorithm we have used in this project is described below:

1. Start
2. Extract Raw Dataset using Twitter API
3. Convert response json to dataset selecting required features
4. Create dictionary of date and stock price using yahoo finance API
5. Apply text cleaning methods to the text
6. Create Sentiment Feature by using pre-trained 'flair' Model
7. Encode the Sentiment Feature
8. Split data into 70% training set and 30% testing set
9. Create, Train and Evaluate different Regression Models
10. Compare and Choose Best Model
11. Predict Next Stock Closing Price Using Best Model
12. Apply Abstractive Summarization

13. End

o Explanation of implementation

First of all, we had extracted raw data in the form of json from Twitter API. Using the extracted json file, we had created pandas dataframe out of it by selecting only the required features such as tweet text, friend's count, retweet count, follower's count and so on. In order to create a label for each tweet, we had created a dictionary of date and stock price by calling yahoo finance API and mapped each tweet row with the stock price in the dataframe. After that, various text cleaning methods had been applied such as cleaning special character and punctuation, performing stemming and lemmatization, converting to lowercase, removing meaningless words and removing stop words. Also, using the tweet text in the dataframe, sentiment feature was created by using a pre-trained flair module and was encoded further to 1 for POSITIVE and -1 for NEGATIVE sentiment. After performing all of this dataset creation and feature engineering, the output dataset had been passed for further processing.

In the next step, we splitted the dataset into 70% training set and 30 % testing set. Then, we used four different models to find the best model for this analysis. Those models are LSTM, Linear Regression, Decision Tree Regression and Knn Regression. After running all of the models, we have calculated the root mean square error (RMSE) and the best model was found to be Linear Regression as shown in the figure below.

	Model	RMSE
0	LSTM	838.239920
1	Linear Regression	49.487728
2	Decision Tree Regression	129.527437
3	KNN Regression	68.506264

Fig: Model Comparison using RMSE score

After getting the best model i.e. Linear Regression, we called the Twitter API to get the latest sentiment and passed it through all the process again to create features out of it and passed through a trained Linear Regression model to predict the next closing price. Finally, we used the pre-trained abstractive summarization model by passing the latest sentiments to get an abstractive summary.

3.7 Results

o Diagrams for results with detailed explanation

We ran our linear regression model on 23rd November 2021. The last closing stock price shown by our model is $1156.86, and the next closing stock price predicted by our model is $1099.30.

```
[30]: # Current Date of Model Run
print('Current Date:                    ', dt.date.today())
print('Last Stock Closing Price:        ', y_test.iloc[-1])

# Next day price using Linear Regression
print('Next Stock Closing Price:        ', next_closing_price)

Current Date:                    2021-11-23
Last Stock Closing Price:        1156.8699951171875
Next Stock Closing Price:        1099.3018928804145
```

Fig: Stock Price Prediction

We called the Twitter API to get the current sentiment of the users. Our model took the last four/five tweets and produced the following paragraph. Then, we used this paragraph as an input in our abstractive summarization model.

```
"@teslaownersSV @elonmusk Yeah but I heard they’re going to do the Tesla bot in Palo Alto and it will clean our things and make croissants and coffee ☕ in the morning 🙏RT @TravisAllen02: Massie has an engineer
ing degree from MIT, has solar panels on his house, drives a Tesla, and has convinced rural Kentuc…RT @invest_answers: IA #TATuesdays – the best all-around indicators all wrapped up in a simple easy to learn vi
deo – #Bitcoin market has a…No new Form 4's from Tesla yesterday or today yet. That's good news for the split rumor!\n$TSLA #120921 #TESLART @TravisAllen02: Massie has an engineering degree from MIT, has solar
panels on his house, drives a Tesla, and has convinced rural Kentuc…RT @silverph: @elonmusk “Dad, the Tesla’s talking!” 😂 Tobie is in the spotlight! I wish I can open and close the frunk automatically to giv…R
T @TravisAllen02: Massie has an engineering degree from MIT, has solar panels on his house, drives a Tesla, and has convinced rural Kentuc…@sirbu_ion1 @crypto_birb @elonmusk @Tesla I wasn’t arrogant btw., we lc
ve birb 🥰RT @creepeffect: The tesla tunnel thing is funny enough in its own right but it’s even funnier when you consider that the idea exists becau…RT @bear_wrongdoer: Me and my friends waiting for out turn o
n the tesla hyperloop https://t.co/OjGeLaZgkiRanjan, G., Rao, A.S.R.: Basic and applied soil mechanics. New Age Int. (2007)\n@RayDalio @OracleResearch @federalreserve @USUN @UniofOxford @Tesla @elonmusk @larrye
llisonNusier, O.K., Alawneh, A.S.: Micropile technique to control upward movement of lightweight structures over expansive soils. Geotech. Geol. Eng. 22(1), 89 (2004)\n@RayDalio @OracleResearch @federalreserve
@USUN @UniofOxford @Tesla @elonmusk @larryellisonAli, M., Ahmed, S.M.: Micropile technique to control heave on expansive soils. In; Proceedings of Indian Geotechnical Conference, Kochi (paper No. D311) (2011)\r
@RayDalio @OracleResearch @federalreserve @USUN @UniofOxford @Tesla @elonmusk @larryellisonRT @resmiyunusemre: 🌟MAGIC BNB🌟\n💰15% BNB REWARDS \n💰50 #NFTs TO HOLDERS EVERY WEEK \n💰1000$ AIRDROP TO HOLDERS \n💰
CARS GIVE AWAY TO HOLDERS (…For those of you who haven't kept up, including Guatam here apparently, what Boring Company have ACTUALLY built is NOT the tunnels in the video\n\nThey've built a short pedestrian-si
zed tunnel under Las Vegas in which a taxi driver in a Tesla drives you along at 30mph https://t.co/tf24Df15BI"
```

Fig: Twitter Latest Sentiment

After getting the sentiment data, we ran the abstraction model. The model produced the following abstract summary.

Fig:

```
# Displaying Summarized Text
print(summary_text)
```

```
No new Form 4's from Tesla yesterday or today yet yet . That's good news for the split rumor! @teslaownersSV @elonmusk Yeah but I heard they’re going to do the Tesla bot in Palo Alto and it will clean our thin
gs and make croissants and coffee in the morning .
```

Fig: Abstractive Summarization of latest sentiment

3.8 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)
Aashish Pandey	25%	train and evaluate the model
Anjan Shrestha	25%	Create different regression model
Puja Dhungana	25%	apply abstractive summarization
Umesh Jaiswal	25%	generate the final prediction

Issues/Concerns:

During the initial stage of our project, we couldn't determine the time lag between the tweet date and the change in stock. For instance, we could not determine if the tweet from Sunday affects the price of Sunday, Monday, or Tuesday. Now, we have solved this issue. However, due to the strict API, we couldn't collect data from Facebook and Discord. So, our results are based on data collected from Twitter.

References

[1] *S&P 500 stock data*. (n.d.). Kaggle. Retrieved November 27, 2021, from https://www.kaggle.com/camnugent/sandp500

Twitter mood predicts the stock market. (2021, 0ct 14). arxiv.org. https://arxiv.org/pdf/1010.3003&

[2] Kim, M., Park, E. L., & Cho, S. (2018, Dec). Stock price prediction through sentiment analysis of corporate disclosures using distributed representation. *Intelligent Data Analysis 22*, 20. 329769006_Stock_price_prediction_through_sentiment_analysis_of_corporate_disclosures_using_distributed_representation

[3] Bhavsar, K. (n.d.). *Stock Market Prediction with Linear Regression [For Beginner] | Data Science and Machine Learning*. Kaggle. Retrieved November 27, 2021, from https://www.kaggle.com/getting-started/94106

[4] Chung, H., & shik Shin, K. -. (2018, October 28). *Genetic Algorithm-Optimized Long Short-Term Memory Network for Stock Market Prediction*. MDPI. https://www.mdpi.com/2071-1050/10/10/3765/htm

[5] Long, W., Lu, Z., & Cui, L. (2018, october 24). *Deep learning-based feature engineering for stock price movement prediction*. sciencedirect.com. https://www.sciencedirect.com/science/article/abs/pii/S0950705118305264?via%3Dihub

[6] Nguyen, T. H., Shirai, K., & Velcin, J. (2015, December 30). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603-9611. Y5uDZQfiNP9QVCG7bEkwI4RgE3q_eHF6BVGapR4IMtnP_sJ9CvQoE5c7_

[7] Cakra, Y. E., & Trisedya, B. D. (2015). *Stock price prediction using linear regression based on sentiment analysis*. IEEE Xplore. https://ieeexplore-ieee-org.libproxy.library.unt.edu/document/7415179

[8] Mittal, A. (2012, 00 00). *Stock Prediction Using Twitter Sentiment Analysis*. cs229.stanford.edu. https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf