

# Analysis on the effects of wallstreetbets on the stock market

by

Abir Hossain

17201017

Zahid Sarker

18101644

Md Sadman Radib

19101653

Sayed Bin Taher

18341002

Syed Mohebbullah Samrat

18101647

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
January 2022

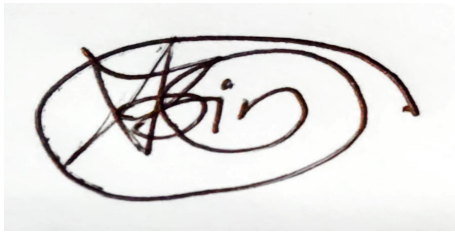
© 2022. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

## Student's Full Name & Signature:



---

Abir Hossain  
17201017



---

Zahid Sarker  
18101644



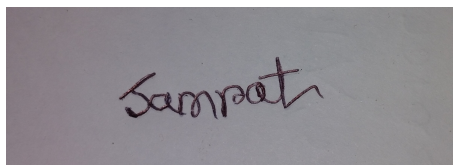
---

Md Sadman Radib  
19101653



---

Sayed Bin Taher  
18341002



---

Syed Mohebbullah Samrat  
18101647

# Approval

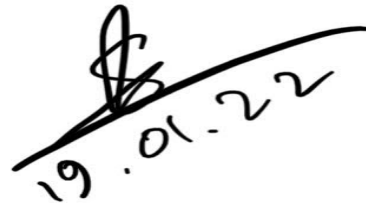
The thesis/project titled “Analysis on the effects of r/wallstreetbets on the stock market” submitted by

1. Abir Hossain(17201017)
2. Zahid Sarker(18101644)
3. Md Sadman Radib(19101653)
4. Sayed Bin Taher(18341002)
5. Syed Mohebbullah Samrat(18101647)

Of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 20, 2022.

## Examining Committee:

Supervisor:  
(Member)



19.01.22

---

Mr. Arif Shakil  
Lecturer  
Department of Computer Science and Engineering  
Brac University

Supervisor:  
(Member)



---

Mr. A. M. Esfar-E-Alam  
Lecturer  
Department of Computer Science and Engineering  
Brac University

Co-Supervisor:  
(Member)



---

Mr. Tanvir Rahman  
Lecturer  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

Sadia Hamid Kazi, PhD  
Chairperson and Associate Professor  
Department of Computer Science and Engineering  
Brac University

## Abstract

In January of 2021, the stocks of various companies, i.e. Gamestop, AMC etc. had a huge surge in price. The incident was attributed to redditors of the subreddit called r/wallstreetbets making a massively coordinated strategy to short squeeze the market. It is likely that the stocks prices were directly affected by the aforementioned subreddit. To better understand the effect on the market wallstreetbets had, we will analyse the reddit posts from wallstreetbets subreddit with stock prices data set. Our intention is to understand this phenomena and find out if the subreddit really had a substantial effect on the market. We also like run our model to figure the level of involvement with the stocks of GME and AMC In our research, we show that r/Wallstreetbets had a direct influence and correlation on various stock prices by using sentiment analysis and supervised machine learning algorithms.

**Keywords:** Stock Market, Subreddit, Wallstreetbets, GameStop, AMC, Sentiment Analysis, KNN, SVM, Decision Tree, Naive Bayes.

## **Dedication**

This work of ours is dedicated to our parents, who were with us in every step of our lives. We could not have reached this far in our lives without their ever present blessings and support. And also special thanks to our supervisors who provided us their utmost support.

## **Acknowledgement**

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisors Mr. Arif Shakil sir, Mr. A. M. Esfar-E-Alam sir, and Mr. Tanvir Rahman sir for their kind support and advice in our work and who helped us whenever we needed help.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

# Table of Contents

Declaration	i
Approval	ii
Abstract	iv
Dedication	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	ix
List of Tables	x
Nomenclature	xi
<b>1 Introduction</b>	<b>1</b>
<b>2 Background Information</b>	<b>3</b>
2.1 Research Problem . . . . .	3
2.2 Research Objectives . . . . .	4
2.3 Literature Review . . . . .	5
2.3.1 Literature Search Process . . . . .	5
2.3.2 Related Works . . . . .	5
<b>3 Dataset Preparation and Methodology</b>	<b>8</b>
3.1 Workplan . . . . .	8
3.2 Input Data . . . . .	10
3.3 Data Preprocessing . . . . .	10
3.4 Used Architecture . . . . .	12
<b>4 Implementation and Result Analysis</b>	<b>14</b>
4.1 Implementations . . . . .	14
4.2 Results . . . . .	16
<b>5 Conclusion</b>	<b>33</b>
5.1 Future Works . . . . .	33





# List of Figures

2.1	Literature Search Process . . . . .	5
3.1	Workflow for achieving research objectives . . . . .	9
4.1	Proposed model . . . . .	15
4.2	Word Cloud of GME related posts . . . . .	16
4.3	Word Cloud of AMC related posts . . . . .	17
4.4	Historical price of GME in 2021 . . . . .	17
4.5	Historical price of AMC in 2021 . . . . .	18
4.6	Prices of both stocks in one frame . . . . .	18
4.7	Volume of GME . . . . .	20
4.8	Volume of AMC . . . . .	20
4.9	Volume of both stocks in one frame . . . . .	21
4.10	Total volume, score and comments for GME and AMC . . . . .	22
4.11	Relation of GME prices with polarity . . . . .	22
4.12	Relation of AMC prices with polarity . . . . .	23
4.13	GME price relation with comment . . . . .	23
4.14	AMC price relation with comment . . . . .	24
4.15	Price relation with Score for GME . . . . .	24
4.16	Price relation with Score for AMC . . . . .	25
4.17	Polarity vs Score graph GME . . . . .	25
4.18	Polarity vs Score graph AMC . . . . .	26
4.19	GME polarity vs total comments . . . . .	26
4.20	AMC polarity vs total comments . . . . .	27
4.21	Correlation of volume and polarity of GME . . . . .	28
4.22	Correlation of volume and polarity of AMC . . . . .	28
4.23	GME price and polarity in one minute interval . . . . .	32
4.24	AMC price and polarity in one minute interval . . . . .	32

# List of Tables

4.1	Number of posts, comments score for both stocks . . . . .	19
4.2	Volume of stocks involved for both GME and AMC . . . . .	21
4.3	Performance for 90 days GME . . . . .	29
4.4	Performance for 60 days GME . . . . .	29
4.5	Performance for 30 days GME . . . . .	29
4.6	Performance for 90 days AMC . . . . .	30
4.7	Performance for 60 days AMC . . . . .	30
4.8	Performance for 30 days AMC . . . . .	30

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

*AMC* refers to AMC Entertainment Holdings, Inc. stock

*DT* Decision Tree

*EnsembleLearning* Maximum voting ensemble

*GME* refers to Gamestop Corp. stock

*KNN* k-nearest neighbors algorithm

*ML* Machine Learning

*SVM* Support vector machine

*WSB* r/wallstreetbets subreddit

# Chapter 1

## Introduction

In January 2021, we witnessed Gamestop stocks soar up approximately 600% along with other stocks i.e amc, gme etc., which was an unprecedented incident. As it was later found out that a subreddit r/wallstreetbets was involved in the whole incident as they created an artificial demand for the involved stocks to hike up the prices. This was a very highly coordinated movement which was a new kind of market manipulation. As it was one of the first of its kind open source market manipulation. This kind of manipulation is very new to everyone and causes massive harm to investors which can very well lead to loss of faith in the market. This could be disastrous to the economy. So this brings forward the question how big of a role did WSB play in this situation. So we wish to make a system to determine this involvement and answer this question.

To achieve our goal we require a better understanding of the stock market. The stock is a sort of market where there exists the action of buying, selling or investing on shares of a publicly owned company. People can buy stock of a particular company and also sell them when the price of that stock increases than the buying price. Stock market is one of the important pillars of a country's economy because it gives the company an opportunity to acquire capital from the public. The two main things that the stock market provides is that it gives the companies the capital it needs to enlarge their business and also helps investors make profit by selling the stock when its price increases by a significant amount. Though the market is very versatile it can still be manipulated for personal gain.

Market manipulation is a technique or combination of multiple techniques used by the traders in order to gain advantages in the stock market. It is difficult to define market manipulation based on parameters and attributes because market manipulation takes on many forms. In our paper we define market manipulation simply as gaining advantages in the stock market by using different techniques. These techniques may involve increasing or decreasing the price intentionally to gain personal profit, releasing false information, using computer algorithms to manipulate the market and many more.

There are numerous methods for manipulating the stock market. Social media or telemarketing can be used to manipulate the market. It is done for monetary gain. Market manipulation can be used to purposefully drive the price of a stock up or

down, having a significant impact on the stock market. The price movement benefits the manipulators, while the investors who were enticed in suffer losses or lose their money.

Like we said, social media can play a vital role in manipulating the market. WSB is a subreddit in reddit, a blogging platform akin to other social media platforms, where members mainly brainstorm stocks and options trading. It is mostly known for the members' colorful and vulgar lingo and aggressive trading tactics where in most cases the trading strategies they employ are highly speculative and risky in nature. The community prides itself as an outsider to the trading game the more established market influencers i.e large capital investors and hedge funds companies. They are also notable for their GME short squeeze debacle that incurred huge losses for a few U.S. hedge firms and short sellers in just a couple of days in early 2021. Though GME was mostly attributed to the whole incident, other stocks were also involved , i.e AMC, etc. The entire incident can be categorized as market manipulation.

Fortunately, by using sentiment analysis techniques we can determine such market manipulation by finding the correlations. By taking public sentiment into account we further make a model that can determine the effect of WSB. There have been various works where sentiment analysis was used in the models[23]. We would also like to use the sentiment of WSB on the stock market to make a model and analyze our findings.

# Chapter 2

## Background Information

### 2.1 Research Problem

Market manipulation is a big concern in the stock market where the most concerning issue is that this can affect the stock market tremendously and it is harmful for the other traders who do not have savvy or know-how of the manipulation which will eventually become harmful for the whole stock market[24]. This especially became more concerning with the emergence of r/wallstreetbets into the mainstream. As they highly coordinated manipulation of the stocks of various shorted stocks. This incident was unprecedented as a manipulation of such a large scale had never been seen before where essentially people from the entire world were instigated into a short squeeze of the market. There is a risk of more situations like that occurring in the future. So it is imperative that people know when such a market manipulation is happening on the stock market.

However, the lack of an effective and efficient detection system makes it difficult for regulators or enforcers to track the stock market and detect manipulation in real time. There is a lot of information out there on stock market manipulation hypotheses [8][3][1] and a few empirical examinations of real-world situations of manipulation [4]. However, due to a lack of understanding of strategic spoofing strategies, an effective manipulation detection model is highly sought after. So by using public sentiment to detect manipulation as they happen would be helpful to all involved with the stock market. By effective analysis of the market, investors can take precautions to protect themselves from such incidents in the future.

One of the major problems for finding appropriate data for market manipulation is that there are not enough datasets. It is one of the major constraints to dive deep into the research so that we can find out accurate ways to analyze the stock market. Aggarwal and Wu (2006) got 141 manipulation cases but unfortunately, they can only obtain data for 51 cases from 1900-2001 which is not sufficient[3]. Comerton-Forde and Putniņš [6] used hand collected data to examine manipulation with the use of Detection Controlled Estimation (DCE). Only a tiny amount of manipulation data was examined which is around one in three hundred cases. Incomplete detection and non-randomness of the sample is another limitation of research regarding stock market manipulation.

There is no defined method to detect market manipulation. Analyzing the vast amount of posts from r/wallstreetbets and huge amounts of data from the stock market which keeps increasing is a challenging task. The existing top-down approach for detecting market manipulation based on a predefined pattern has to rely on only the rules set by the experts but due to the exponentially increasing data it is not a very good choice to use predefined patterns[10]. Moreover, the manipulators can easily find new techniques to bypass the predefined rules. Therefore, we use different supervised machine learning algorithms for our research to find the most accurate model. We will take the r/wallstreetbets data to analyze the public sentiments and use that to detect and understand market manipulation using supervised learning algorithms, such as - Naive Bayes, SVM, KNN, RNN, Decision Tree.

## 2.2 Research Objectives

The research aims to make a system to detect market manipulation of the stocks in the stock market with the use of neural networks and machine learning algorithms. The r/wallstreetbets subreddit dataset is fed to a Textblob to determine the sentiment and then it will be used in conjunction with the stock market datasets to make a quick and accurate detector model[23]. The overall objectives of this research are to-

1. Understanding how the stock market behaves and how it is manipulated.
2. Converting raw text data from r/wallstreetbets to a more understandable format.
3. Classify the data to a proper sentiment classification.
4. Developing Naive Bayes, SVM, KNN, RNN, Decision Tree models for analysis.
5. Evaluate the various models and determine the performance using various performance metrics.
6. Compare performances of various models to determine the best performing models.
7. Further improving and tweaking the performance of the models.



## 2.3 Literature Review

### 2.3.1 Literature Search Process

Having a specific research topic, we have to have a structured literature search process. We used Kaggle(<https://www.kaggle.com/>) for finding our datasets by searching "wallstreetbets" and Yahoo! finance for stock data. We also searched for paper with the string "Stock Market" which has a large record list. So, we have to narrow it down by specifying the search string gradually. The search process is shown in fig 3.1 :

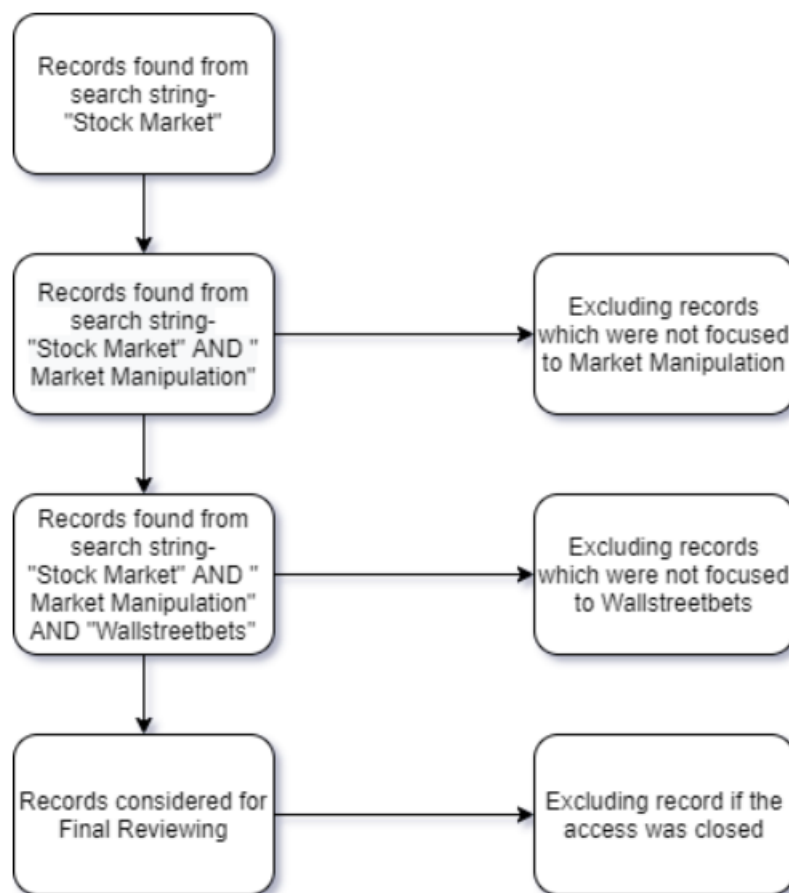


Figure 2.1: Literature Search Process

### 2.3.2 Related Works

Research about stock market prediction is nothing new. There has been a significant amount of research done on market manipulation too. However, predicting and detecting market manipulation has been a challenging task. The existing top-down approach uses predefined rules for detecting manipulation [10] . [10] [22] [25] [16] have used supervised learning algorithms to detect market manipulation. Some research has also been conducted by using sentiment analysis of various social media platforms like Facebook [9], Twitter [10][26] [20] to find the influence of social

media on the market. However, research on the influence of reddit, more specifically r/wallstreetbets subreddit, is fairly new and few in number. Researchers have turned their attention to r/wallstreetbets only after the recent incident of Gamestop frenzy in the beginning of 2021. Research on the influence of r/wallstreetbets on market manipulation has been increasing since, which also motivated us to further research about r/wallstreetbets and market manipulation. [27] have extracted 10.8 million comments from reddit and used sentiment analysis to understand the influence of r/wallstreetbets on GameStop stock prices from January to February 2021. They extracted various emotions using the “Text2emotions” package of Python and divided the emotions into 5 groups - happy, angry, sad, surprise and fear and found that these emotions influence the GameStop stock differently.

We have also found that JB approach and SNLP was used for sentiment analysis by [23]. The author then used the sentiment data and aggregated them by a specific date. They got their sentiment data from twitter. Rao and Srivastava also used twitter to analyze the stock market[5]. The author used WEKA machine learning library as a training algorithm to train their model. For evaluating the performance the author used metrics such as *Accuracy*, *Precision<sub>p</sub>*, *Recall<sub>p</sub>*, *F – measure<sub>p</sub>*. They found SMO was the best performing algorithm for stock market prediction. On the other hand, ASC and Bagging showed the worst performance. Prediction with the highest accuracy of stock market trends could be achieved on day 7 in general from the day of the initial prediction where day 7 was the last day. The algorithms generally attain accuracy in the range of 60-70%[23]. LWL showed a relatively high performance when classifying trends in appropriate classes. They ascertained that that accuracy improvement in prediction was in the range of 0-3% with the aid sentiment analysis. The authors also validated their model using 10 fold cross-validation. They also concluded sentiment lexicons that were available were not providing high enough accuracy. The authors of [16] used supervised learning algorithms to detect market manipulation using supervised learning algorithms, i.e. SVM, LDA, QDA, DTC, ANN, KNN, LR .They used accuracy, sensitivity, specificity to figure out the performance. They detailed that a high sensitivity value for a given specificity value indicated a better performance. They used 5 fold cross-validation. ANN was the best performing in terms of accuracy, sensitivity with a low specificity. But The authors found that supervised learning models have excellent performance for detecting the anomaly in daily trading data but on the other hand, in terms of tick trading data, the performance is much poorer. ANN in 99.9% specificity had a sensitivity of 90.5% for daily trading but in the same specificity it had sensitivity of 3.8%. Attention based NLP i.e Transformers are a relatively new phenomenon. But it shows great promise as it can achieve a higher level of efficiency and speed. Attention based NLP could be implemented to find sentiment of text based data[17].

Somewhat recently, sentiment analysis has acquired significance in light of the accessibility of gigantic measures of printed information on the online media. The printed information that is available in the online sources can be drawn upon to assess clients in different scenarios. This large volume of printed information can be used in for Machine Learning purposes. Thus, Machine Learning scientists have completed exploration on mining assessments of clients of these stages. Twitter is one of the notable platforms that can be used to analyze the sentiment of people.

According to Yuan many researchers investigated dictionary, and machine learning based opinion arrangement strategies. Counting of words as well as focal point methodology were used for reference terminology. Support Vector Machine(SVM), Naïve Bayes(NB) and also Maximum Entropy (ME) algorithms were brought into play for machine learning approach[13]. They inspected BoW(Bag of Words) and Grammatical form phonetic comments having N-Gram highlights. Based on their analysis they concluded that BoW is a powerful and easiest feature that gave the best probable outcome. For numerous application areas, estimation examination of client feelings might be done. SVM, NB and ME machine learning calculation techniques were used by Joshi and Tekchandani to conduct similar examinations and found out that SVM performs better than other algorithms[12]. Like online media, news is additionally a significant outside factor that passes on significant occasions identified with stocks and stock markets, so AI specialists likewise performed assumption examination of information. Dang and Duong performed SVM on news and stock price information to discover the relation between stock price and monetary information. A connection between stock price and news was found in their research[11]. Dang et al. 2018 compared TGRU network with other standard models and found out that the mentioned network surpassed those models with accuracy of 66.32%.[19]. Khare et al. 2017 analyzed momentary future costs of the New York Stock Exchange by using RNN and FFANN[15]. They discovered FFANN's exhibition prevalent in anticipating stock momentary costs. All of the mentioned research methodologies showed promising results in detecting and predicting market manipulation using sentiment analysis and supervised learning algorithms to make our model.

# Chapter 3

## Dataset Preparation and Methodology

### 3.1 Workplan

The main objective of our research is to analyze the influence of r/Wallstreetbets subreddit on different stocks. In order to do the research, we have to extract data from reddit first and then to analyze how r/Wallstreetbets are affecting the various stocks, we need normal and manipulated stock data. After preprocessing the data we need to label the datasets into different train and test sets. We will use the train set to train our model using different supervised learning algorithms and find the best method by testing the accuracy of our model by comparing with the normal and manipulated stock data to detect correlation of its effects.

The entire process can be divided into 3 steps and they are described below -

1. Data pre-processing: In this stage we format the raw data from reddit and we also format the stock data in order to prepare the datasets for further analysis.
2. Sentiment Analysis: In this we run sentiment analysis using textblob and vaderSentiment.
3. Training the model: In this stage we train our model using various supervised machine learning algorithms.
4. Testing the model: In this stage we test the correlation of our model to find the best performing algorithm in order to detect manipulation.

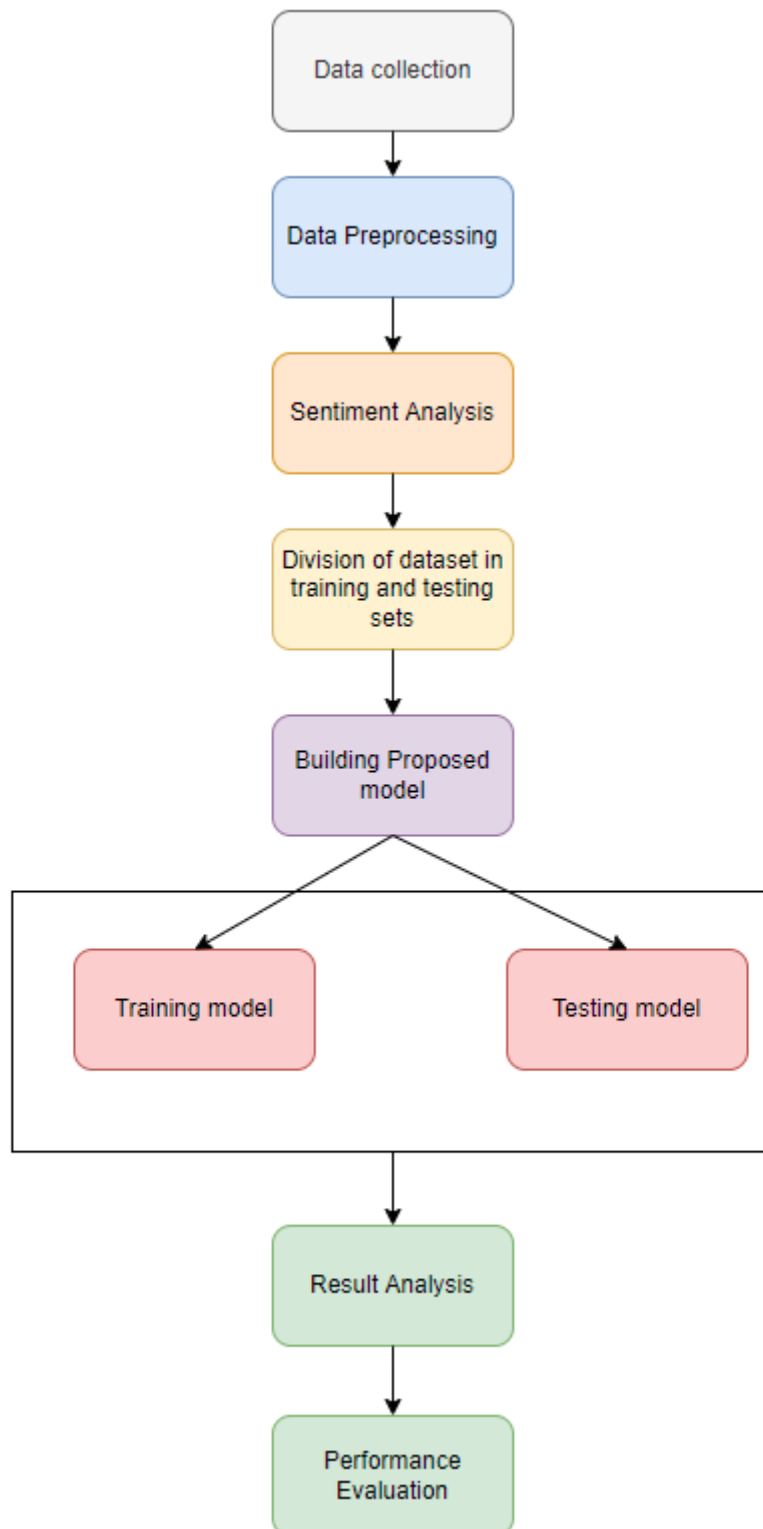


Figure 3.1: Workflow for achieving research objectives

## 3.2 Input Data

There were various datasets on r/wallstreetbets on kaggle. Some focused on the January 2021 incident and others had a more comprehensive dataset. Since we will work on the r/wallstreetbets influence on the stock market in January 2021, we will use a dataset that focuses on that. We will also require stock market prices on the involved stocks. We can acquire such historical data from Yahoo! finance.

1. **Kaggle** : There were various datasets on r/wallstreetbets on kaggle. Some focused on the January 2021 incident and others had a more comprehensive dataset. Since we will work on the r/wallstreetbets influence on the stock market in January 2021, we will use a dataset that focuses on that. Reddit r/wallstreetbets Posts is a dataset made by Gabriel Pedra that is one the most regularly maintained datasets available. We will use this dataset to run our sentiment analysis. This dataset contains posts' data in text format which makes it more usable for us. It has 8 columns including text body and timestamps which will be very useful in our work. Our primary goal is to determine the sentiment of this dataset by running textblob and vader sentiment[7][18]. We find the sentiment of the posts and make a mean average of the sentiment for a certain time period.
2. **Yahoo! finance** : We will also require stock market prices on the involved stocks. Historical financial performance various stocks can be acquired through Yahoo! finance as most researchers use it[23]. We will need the date, opening price, close price, highest price, lowest price, volume[23]. For that purpose we selected a dataset that has data on r/wallstreetbets title, posts, its rating and number of comments with timestamp. We also collected stock data on GME, NOK, AMC. It has data on open, close, high, low prices with volume and timestamp. This dataset will be used in the models.
3. **Barchart API** : Barchart API provides market data with different options. For our research, we needed historical data at 1 minute intervals everyday. For that purpose, we used Barchart's API to collect GME and AMC stock prices at 1 minute intervals everyday.

## 3.3 Data Preprocessing

For the preprocessing we take the following steps.

1. **Data Cleaning** : Firstly, we take the r/wallstreetbets dataset and change the timestamp to datetime format to make it easier to read for our code and remove the time data to keep only the date. After that we drop the post with no bodies in them and reindex the posts. From the posts, all unnecessary objects are removed by dropping handlers, urls, special characters, single characters and multiple spaces.
2. **Data Transformation** : We add the title and body of the posts together to make a single posts column for the dataset. Also posts containing the following strings- gme, gamestop, amc, nokia, nok were filtered out.

3. **Data Reduction :** We keep the 'Date', 'post', 'score', 'comms\_num' columns and remove the rest from our processed dataset.

After all preprocessing we get a dataset of (11556, 4) dimensionality.

### 3.4 Used Architecture

One of the python libraries that is used for Natural Language Processing (NLP) is Text Blob, with the use of Natural Language ToolKit (NLTK) text blob can perform its task effectively. NLTK library helps users by providing access to many lexical assets and also helps users to classify and categorize data. Text Blob is very simple but effective which can perform complex analysis and a mass amount of textual data. The return values of a TextBlob are the Polarity and the Subjectivity of a phrase or a sentence. The range of value of polarity is from +1.0 to -1.0, negative sentiment is defined by the value -1.0 and the positive sentiment value is defined by +1.0. Subjectivity also has a range from 0.0 to 1.0, where the value 0.0 means that the sentence is very objective and if the value is 1.0 then the sentence is very subjective.[18]

VADER is a sentiment analyzer tool to calculate the sentiment of text posted on social media. VADER sentiment has negative and positive polarities. It is used to calculate the amount of positive or negative emotion expressed within the text and also calculate the intensity of the emotion. One of the major advantages of VADER sentiment is that it does not need any training data and also it can easily understand the emotion of the text consisting of emojis, cursed words, punctuations, etc. VADER sentiment can be used in any sort of social media text. VADER sentiment gives us the score of positivity of a sentence, the negativity of a sentence, the neutral percentage of the sentence, and lastly the compound score. The compound score is the sum of the three percentages and then scaled between the range -1.0 and +1.0, where -1.0 means most negative and +1.0 means most positive.[7]

Naive Bayes is proved to be the best machine learning algorithm for its heigh capability in text classification by researchers. It is very popular for its scalability. Thus being perfect for large datasets[23].Naive Bayes theorem is a quantitative analysis which uses probability to foretell class members. Due to the usage of class conditional independence, it is relatively fast to compute. Class conditional independence considers all elements of the class does not depend on other elements. We found the usage of Naïve Bayes theorem in [23][14][22].

KNN is a classification algorithm that uses samples as parameter units. In this technique every instances are linked to points in n-dimensional space  $R^N$ . When large training sets are used, the approach is labor demanding, and it did not become popular until higher computational power became available.KNN calculations were shown in articles [16][22].

SVM is an optimal method-based classifier that conducts classifies tasks by creating hyperplanes in a multidimensional space to separate cases with distinct class labels. The SVM approach creates a separated hyperplane which optimum and the margin is maximized between two groups.[16] It has been shown to be effective at classifying tasks with high performance in generalization.The application of SVM is discussed in articles [16][22].



Decision Tree is a very popular algorithm. An empirical tree is a data segmentation generated via the application of a set of simple rules. Depending on the value of one input, an observation is assigned to a segment by each rule. Most researchers have utilized this methodology to find little or large data structures and anticipate their value because of its simplicity and comprehensibility. The results are easy to understand as it can be directly be converted to IF-THEN rules[16]. This is a simple and easy to use algorithm.

Random forest is a popular and easier algorithm that is being used in machine learning. It gives accurate results routinely. The biggest advantage of the random forest algorithm is it is very useful for regression and classification which contributes to the major purposes of machine learning. The random forest algorithm is comprised of many decision trees which are trained by the process of bagging or bootstrap aggregating.

Maximum voting ensemble is a model which utilizes multiple models in conjunction and each model has a single vote. They do a vote pooling and determine the result. The answer that gets voted the most is selected as the result for this model[21].

# Chapter 4

## Implementation and Result Analysis

In this section, we go into implementation of model and the preliminary results we attain from the model. We used python for our code using colab to train and test our model. We show our own implementation and the attained results.

### 4.1 Implementations

In this section we will go into the our implementation of our model.

1. Creating the dataset: After collecting the r/wallstreetbets dataset from Kaggle, we divided the dataset for GME and AMC stock based on filtering 'gme|gamestop' and 'amc' words respectively. Then we merged the dataset containing the GME posts with the gme stock prices dataset and the dataset containing AMC posts with amc stock prices dataset based on date. For creating 1 minute interval stock price data we used the Barchart API to acquire 1 minute interval stock prices of GME and AMC stock and merged the dataset containing the GME posts with the 1-minute interval gme stock prices dataset and the dataset containing AMC posts with 1-minute interval amc stock prices dataset
2. Sentiment Analysis: We have used Python's TextBlob library to find the polarity and subjectivity of the r/wallstreetbets posts. For daily stock price analysis, we took the average sentiment of all the posts of that particular day and merged them with the daily stock prices dataset based on the date. For 1-minute interval stock price analysis, we took the average sentiment of the posts every minute and merged them with the stock prices dataset based on every minute.
3. Stock price analysis based on the sentiment: Finally, we have used different machine learning algorithms (KNN, SVM, Naive Bayes, Decision tree and Random forest) and voting ensemble to determine if the algorithms find any relation between the stock prices and the sentiments on a 30-days interval, 60-days interval and 90-days interval.

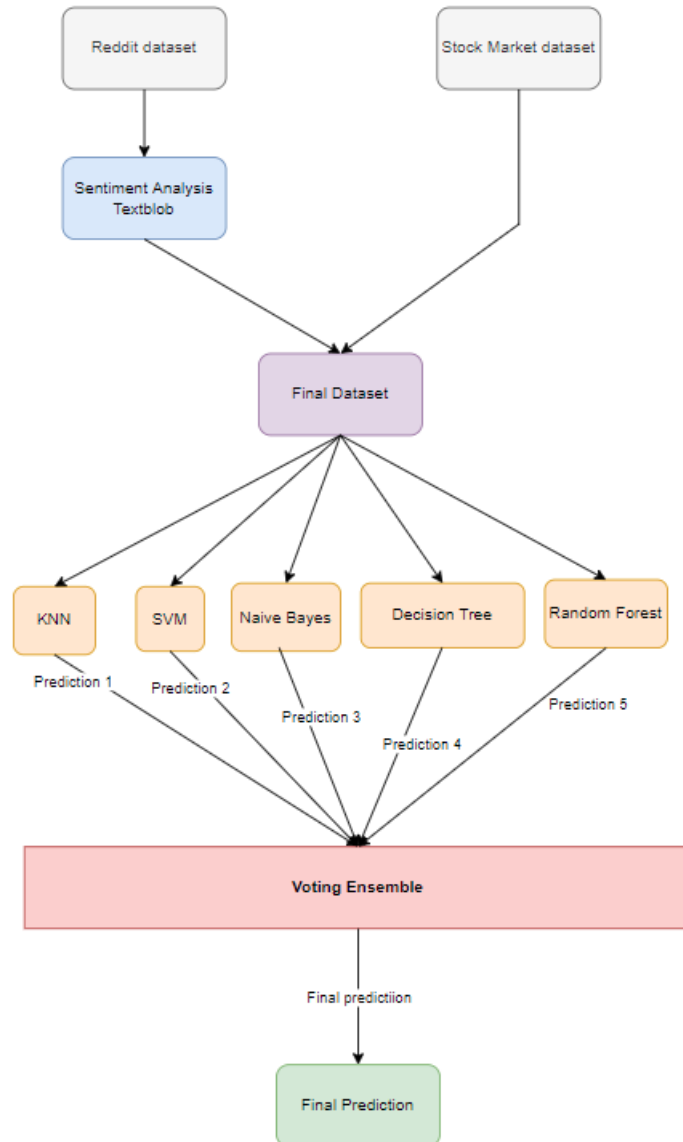


Figure 4.1: Proposed model





the price goes up and down. The GME stock is very volatile and sudden.



Figure 4.5: Historical price of AMC in 2021

In the figure 4.5 we can see the historical data on the price of AMC stock. We also see how the stock behaves. As seen there during the beginning of January the price is stable and it goes on to increase as time goes on. It reaches the peak during the affected time period. The price is unstable during that time as seen through how the price goes up and down. However AMC stock is not as volatile as AMC. The movement of AMC stocks is more gradual.

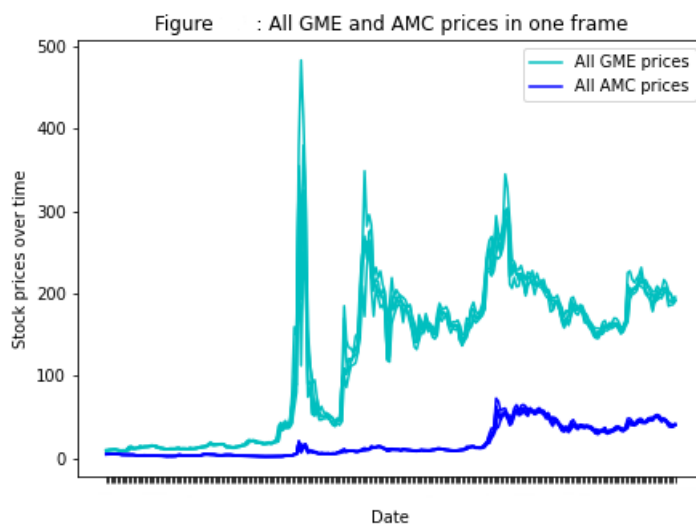


Figure 4.6: Prices of both stocks in one frame

In this figure 4.6 we can see both stocks in one frame. GME has seen a significantly higher increase than AMC as its stock was more affected by WSB. It can be seen in table 4.1 as well. In the table we can see that GME is more talked about in posts and comments of WSB subreddit. The rating of the GME related posts and comments is also higher. So those posts saw a clearly larger level of interaction. The involvement between GME and the WSB is significant. On the other hand, AMC saw less post comments and scores in general. So the interaction in the AMC related topic is significantly lower. So the price movement is also lower.

Stock	Total number of posts	Total no of comments	Total score of the posts
GME	10046	3029284	8820771
AMC	3627	380639	1574459

Table 4.1: Number of posts, comments score for both stocks

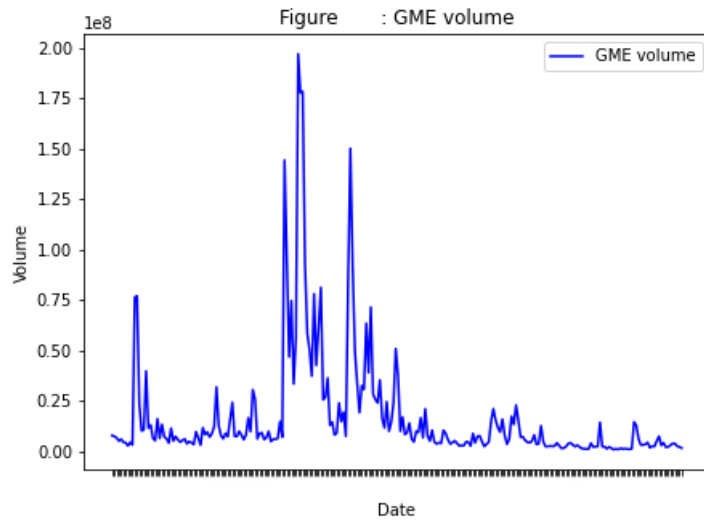


Figure 4.7: Volume of GME

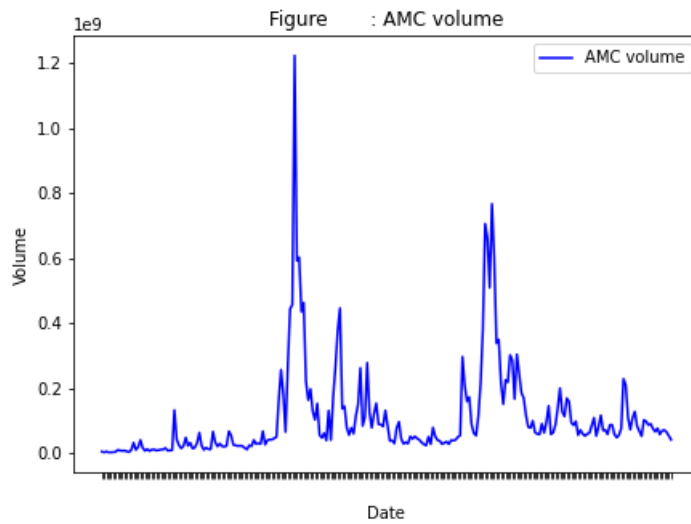


Figure 4.8: Voume of AMC

In the figure 4.7 we can see the historical data on the volume of GME stock. We come to see the stock movement. As seen there during the beginning of January the movement is on the lower end of things but there was an anomaly where the movement spiked. The movement spikes during the affected time period. The movement tapers off quickly after the affected period. In figure 4.8 we can see the historical data on the volume of AMC stock. We see the stock movement. As seen there during the beginning of January the movement is on the lower end of things. The movement spikes during the affected time period. The movement tapers off quickly after the affected period but there is one anomaly not seen in GME is that it sees another spike after the affected period.



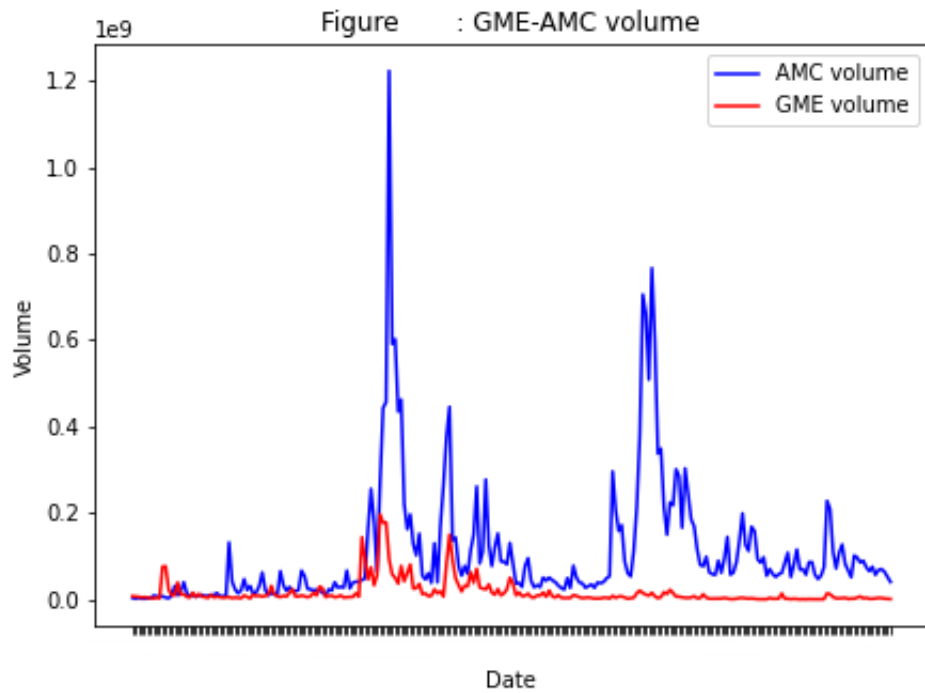


Figure 4.9: Volume of both stocks in one frame

Stock	Volume
GME	4259214700
AMC	27315642900

Table 4.2: Volume of stocks involved for both GME and AMC

In figure 4.9 we can see both the volume of GME and AMC in one frame. AMC has larger volume because the volume of AMC stock is larger than GME. This does not indicate the movement of the stock rather the number of stock between the two. So It can be understood that GME having less stock and more comments and interaction as seen in figure 4.10 that GME is more volatile and the price change is higher due to the lack of stocks in the market. But AMC on the other hand has more stock available and so it has lower price change due to the lower number of comments and interaction with AMC related posts.

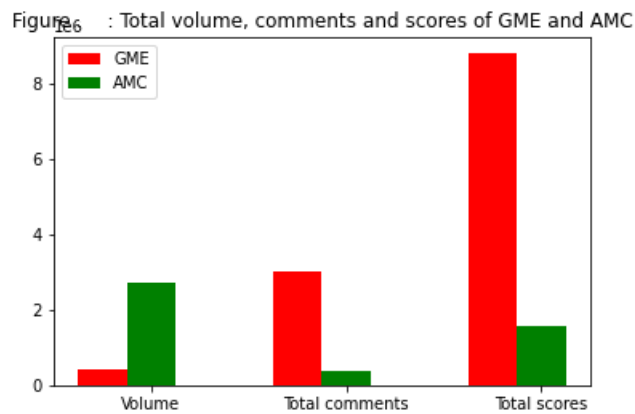


Figure 4.10: Total volume, score and comments for GME and AMC

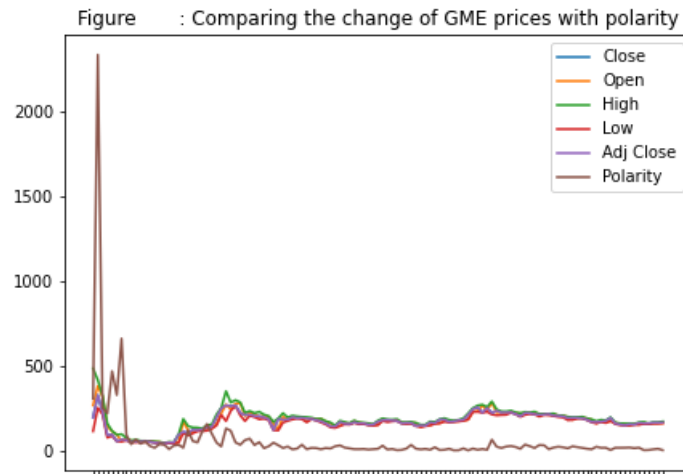


Figure 4.11: Relation of GME prices with polarity

In Figure 4.11 we can see the correlation between the stock prices and polarity. Polarity being higher indicates that the people were more positive about the GME stock prices in reddit. We can see the prices are at its peak, when polarity is high and price is low, when polarity is down. The price graph follows along with the polarity. A pattern is seen there. The pattern indicates the correlation between prices of GME and polarity we attained from sentiment analysis. Thus people being high on GME stock is influencing its price to be higher and vice versa.

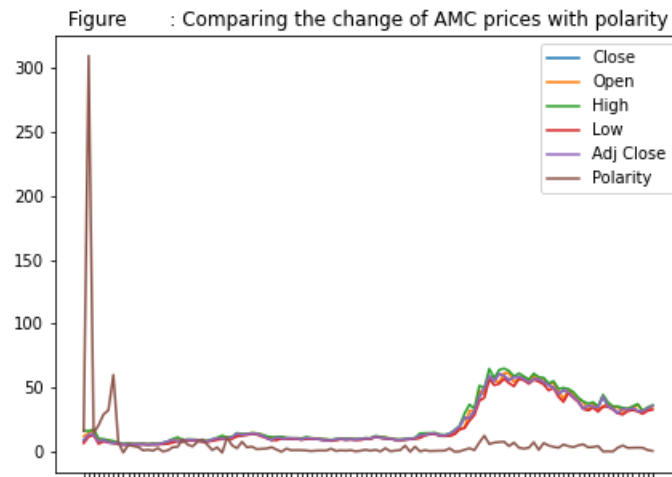


Figure 4.12: Relation of AMC prices with polarity

Figure 4.12 also shows the similar behavioral pattern like in figure 4.11 but with one key difference. In the affected time period the pattern of correlation is similar to GME. As for the difference is that we see the price of AMC increasing in the last end of the graph where the polarity has tapered off due to less post and comment being made. This event is unrelated to the WSB. So we can say that people being high on AMC stock is influencing its price to increase and the opposite case is true as well.

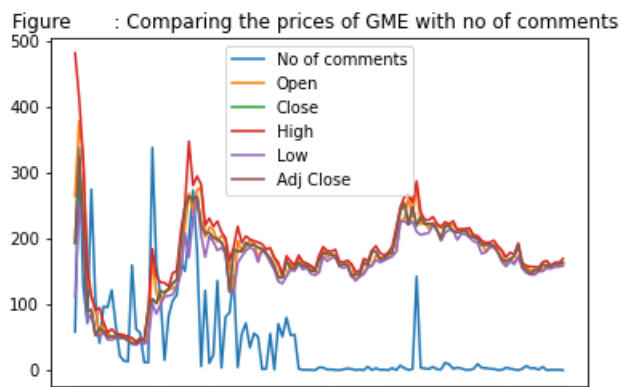


Figure 4.13: GME price relation with comment

Figure 4.13 is a showing correlation between GME prices and number of comments. As the number of comments increase the price increases and as the number of comments decrease the price decreases. However there are some discrepancies at certain points but they are negligible as the pattern between the two follows throughout the graph. The correlation between the two can be understood from this.

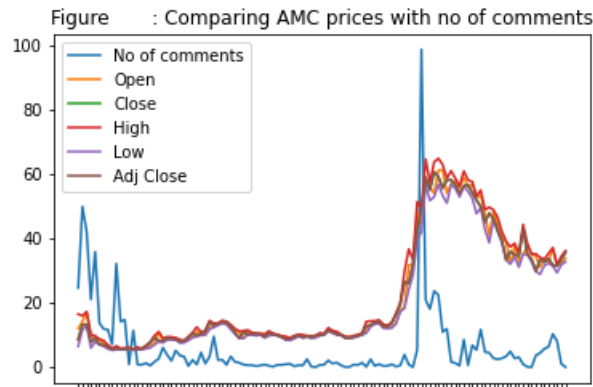


Figure 4.14: AMC price relation with comment

Figure 4.14 shows a similar pattern as seen in figure 4.13, as we can see identical behaviour for AMC and GME stocks. The correlation between AMC prices and number of comments is seen from the graph.

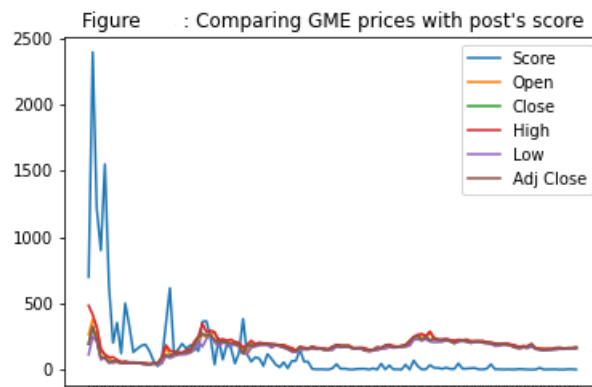


Figure 4.15: Price relation with Score for GME

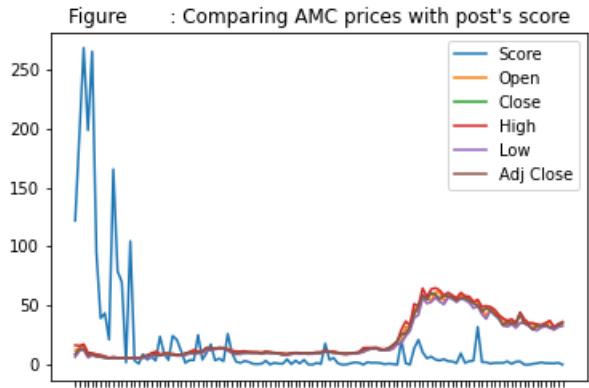


Figure 4.16: Price relation with Score for AMC

From figure 4.15 and 4.16 we can see the pattern prices follow a similar pattern scores of posts. But for AMC there is an anomaly as the price increases when the score tapers off which indicates that this is a different incident as it happened after the stock market short squeeze incident. Overall the correlation between the two stocks is similar.

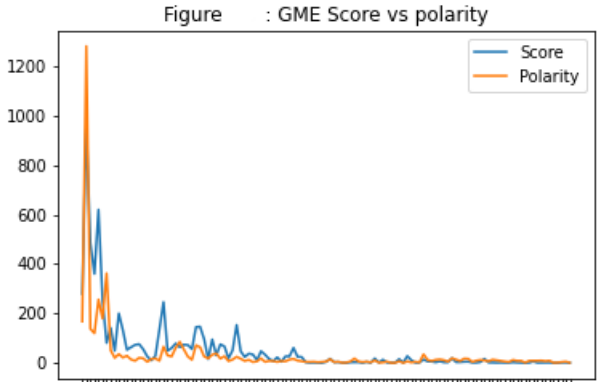


Figure 4.17: Polarity vs Score graph GME

Figure 4.17 and 4.18 indicate that score of posts and polarity have a correlation between each other for both GME and AMC related posts. As polarity follows a similar pattern to scores which indicate interaction in the posts and comments. So we can say polarity and scores have a strong correlation.

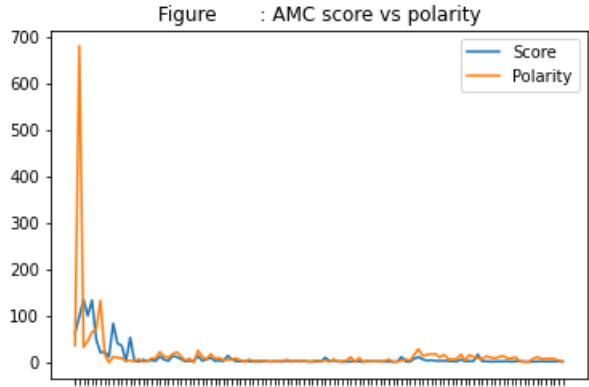


Figure 4.18: Polarity vs Score graph AMC

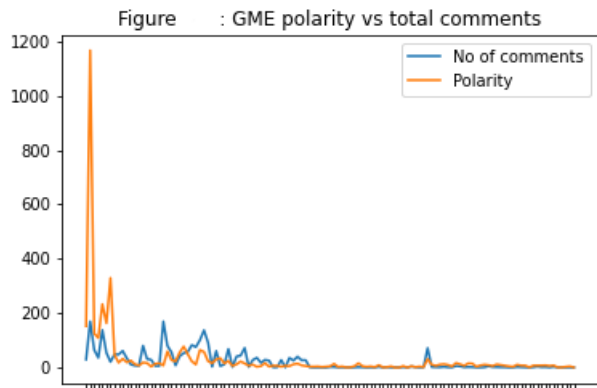


Figure 4.19: GME polarity vs total comments

From figure 4.19 we see the relationship between polarity and number of comments for GME. The correlation between them is clearly visible as the graph shows that both follow a similar trend. The similarity between the two is most pronounced during the affected period of January and February.

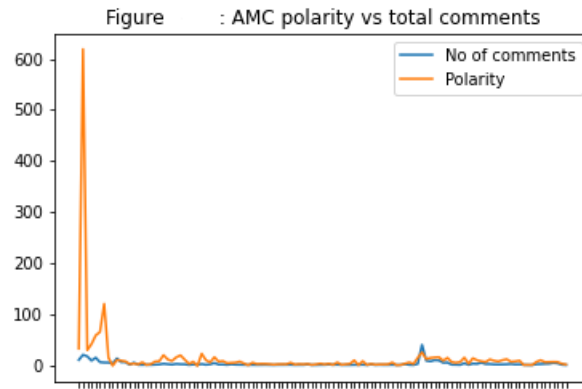


Figure 4.20: AMC polarity vs total comments

From figure 4.20 we see the relationship between polarity and number of comments for AMC. The correlation between them is similar to the GME graph as they have similar trends. Though the number of comments is lower in volume than for GME and so the spikes are less pronounced for the AMC graph.

From figure 4.21 and 4.22 we get the correlation volume of stock and polarity graph for GME and AMC in order. The graphs show there is a spike in the volume of stocks as polarity increases and vice versa for both stocks. It is showing that the positive or negative impression of people regarding GME and AMC is influencing movement of both stocks. The correlation is visible.

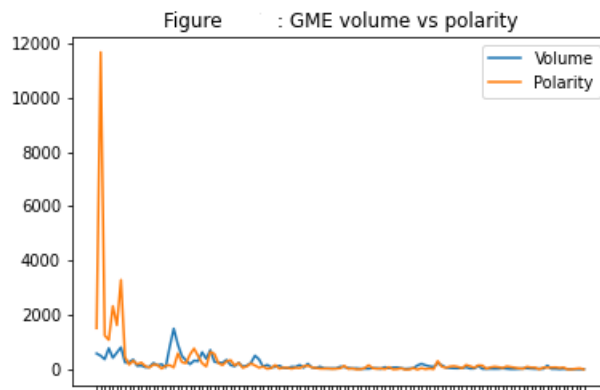


Figure 4.21: Correlation of volume and polarity of GME

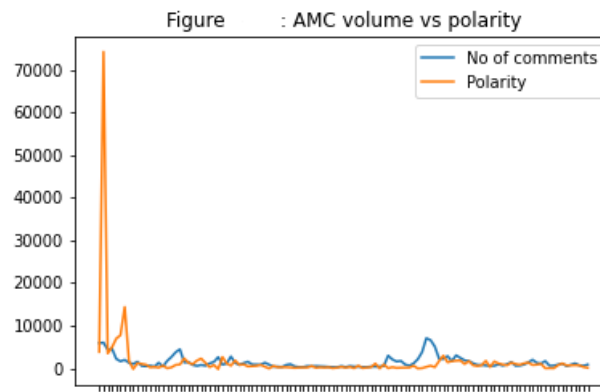


Figure 4.22: Correlation of volume and polarity of AMC



We are showing the model performance with different algorithms and our ensemble model.

Model	Accuracy(%)
KNN	58
Naive Bayes	47.3684
SVM	42
Logistic Regression	47.3682
Random Forest	68
Decision Tree	57.894
Voting Ensemble	57.89

Table 4.3: Performance for 90 days GME

Model	Accuracy(%)
KNN	62
Naive Bayes	53.84615
SVM	69
Logistic Regression	69.2307
Random Forest	69
Decision Tree	61.538
Voting Ensemble	69.2307

Table 4.4: Performance for 60 days GME

Model	Accuracy(%)
KNN	86
Naive Bayes	57.1428571
SVM	86
Logistic Regression	71.428571
Random Forest	86
Decision Tree	71.4285
Voting Ensemble	85.7142

Table 4.5: Performance for 30 days GME

From the tables we see the results we acquire from our used models for GME stock. As seen here the accuracy of our majority voting ensemble model we find that, the accuracy is lowest for the results for 90 days which is 57.89%. If we lower the days used for the model to 60 days. The model accuracy increases to 69.23%. And just for the affected 30 day period we reach highest accuracy performance with 85.71%. From this we can conclude that we get our best result for the most affected 30 days model performance. That performance diminishes as the timespan is increased. Since in the affected 30 days the interaction, post and comments are highest for GME, performance is better. We also come to find that the subreddit posts have a very clear influence on the market for GME stocks. This suggest that the market

behaviour was dictated by WSB on some level.

Model	Accuracy(%)
KNN	63
Naive Bayes	47.3684
SVM	42.1052
Logistic Regression	57.89
Random Forest	68
Decision Tree	63.1578
Voting Ensemble	68.42

Table 4.6: Performance for 90 days AMC

Model	Accuracy(%)
KNN	58
Naive Bayes	47.3684
SVM	42
Logistic Regression	47.3684
Random Forest	68
Decision Tree	57.8947
Voting Ensemble	57.7498

Table 4.7: Performance for 60 days AMC

Model	Accuracy(%)
KNN	71
Naive Bayes	42.8571
SVM	57
Logistic Regression	57.1428
Random Forest	86
Decision Tree	85.7142
Voting Ensemble	85.6031

Table 4.8: Performance for 30 days AMC

From the table we see the results we acquire from our used models for AMC stock. As seen here the accuracy of our majority voting ensemble model we find that, the accuracy of the results for 90 days which is 68.42%. If we lower the days used for the model to 60 days. The model accuracy decreases to 57.89%. This is an anomaly that is unique to the AMC stock. And just for the affected 30 day period we reach highest accuracy performance with 85.71%. From this we can conclude that we get our best result for the most affected 30 days model performance. That performance diminishes as the timespan is increased. Since in the affected 30 days the interaction, post and comments are highest for AMC, performance is better. We also come to find that the subreddit posts have a very clear influence on the market for AMC stocks. This suggest that the market behaviour was dictated by WSB on some level.

To summarize, the prediction results from the machine learning models might not reflect the typical accuracy and we are not concerned about the accuracy because there are other external factors responsible for the stock market prices other than the r/wallstreetbets sentiment but we are only concerned about the r/wallstreetbets sentiment. We used the machine learning models only to see how they perform based on the pattern they find in the sentiment analysis. For example, the models perform well to predict the first 30 days or so because there were a lot of activities at that period and there is a direct correlation between the r/wallstreetbets sentiment and stock prices. As more time passes, the r/wallstreetbets sentiment becomes flat and the accuracy of the model decreases respectively for 60 days and 90 days period. Therefore, to provide more concrete ideas we go deeper by analyzing the stock prices based on intraday returns. Instead of giving a daily summary, we analyze the stock prices for 1 minute intervals everyday. Figure 4.23 and figure 4.24 depicts the GME and AMC stock prices respectively in 1 minute interval.

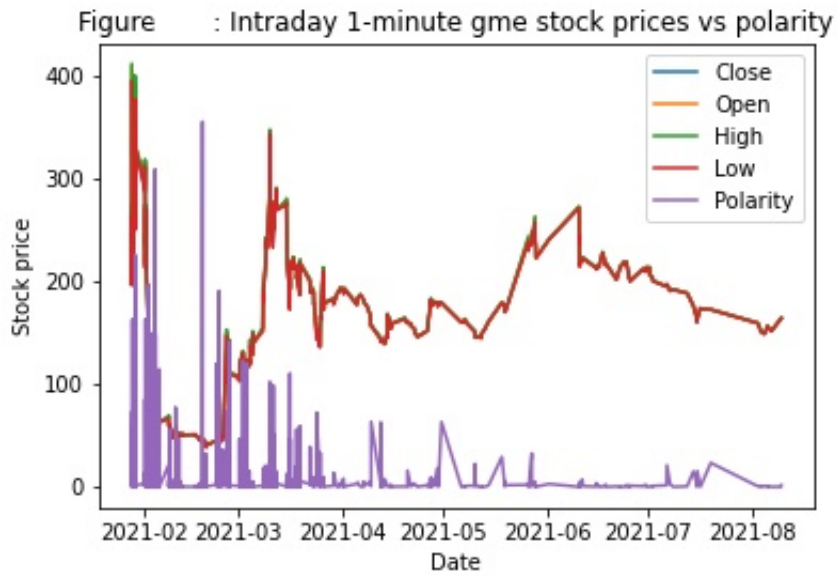


Figure 4.23: GME price and polarity in one minute interval

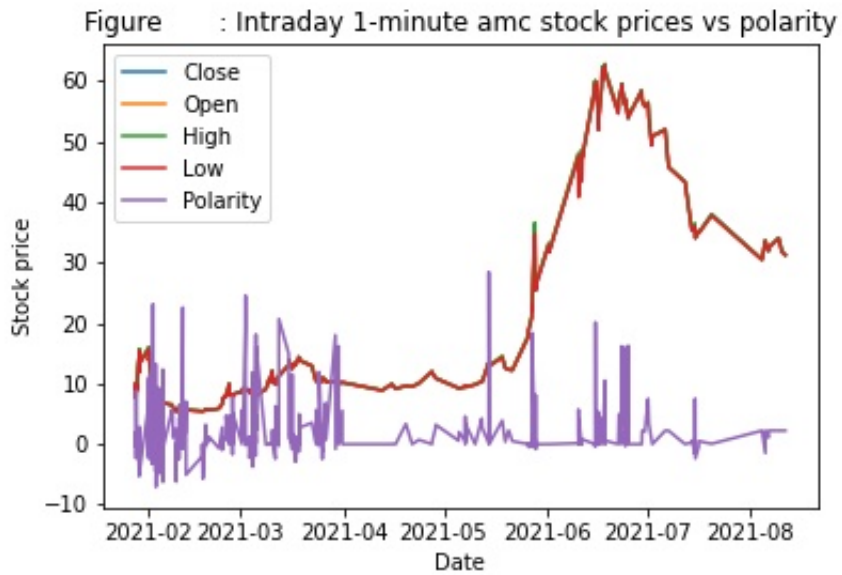


Figure 4.24: AMC price and polarity in one minute interval

# Chapter 5

## Conclusion

Market manipulation with the aid of social media while is a relatively new concept, it is not something to scoff at. As the January 2021 r/wallstreetbets incident can attest to that where that community effectively short squeezed the wall street hedge fund managers. While situations like this happens few and far between, still it can have a catastrophic effect on the market and even collapse the economy[2]. Our work find a significant relation between stock market and WSB. Our model result suggests that the during the time period of the short squeeze, the involvement of WSB has played a significant role in behaviour of GME and AMC stocks due to the higher level of accuracy we are achieving during that time. As the time period is increased the post and comments relating to GME and AMC decreases thus becoming harder to predict the market behaviour. So this proves our hypothesis that the stock market was influenced by WSB during the short squeeze period in 2021.

### 5.1 Future Works

In future we would like to further delve on this topic with a more broader scope. As we would like to work with WSB post and comment on more stocks and the influence it plays. As we only worked with two stocks this time. Our hope is to see if there are any stock that might have been influenced by WSB but was not reported on or noticed by others. We also hope to research comparative behaviour of stock that were and were not influenced during the short squeeze time period. We would also like use a different methods of sentiment analysis. Transformer based models are very intriguing for our research.

# Bibliography

- [1] F. Allen and G. Gorton, “Stock price manipulation, market microstructure and asymmetric information,” *European Economic Review*, vol. 36, no. 2-3, pp. 624–630, 1992. [Online]. Available: <https://EconPapers.repec.org/RePEc:eee:eecrev:v:36:y:1992:i:2-3:p:624-630>.
- [2] F. Mishkin and E. White, “U.s. stock market crashes and their aftermath: Implications for monetary policy,” Jul. 2002. [Online]. Available: [https://www.researchgate.net/publication/5196964\\_US\\_Stock\\_Market\\_Crashes\\_and\\_Their\\_Aftermath\\_Implications\\_For\\_Monetary\\_Policy](https://www.researchgate.net/publication/5196964_US_Stock_Market_Crashes_and_Their_Aftermath_Implications_For_Monetary_Policy).
- [3] R. Aggarwal and G. Wu, “Stock market manipulations,” *The Journal of Business*, vol. 79, no. 4, pp. 1915–1954, 2006. [Online]. Available: <https://EconPapers.repec.org/RePEc:ucp:jnlbus:v:79:y:2006:i:4:p:1915-1954>.
- [4] E. J. Lee, K. S. Eom, and K. S. Park, “Microstructure-based manipulation: Strategic behavior and performance of spoofing traders,” *Journal of Financial Markets*, vol. 16, no. 2, pp. 227–252, 2013, ISSN: 1386-4181. DOI: <https://doi.org/10.1016/j.finmar.2012.05.004>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1386418112000377>.
- [5] T. Rao and S. Srivastava, “Analyzing stock market movements using twitter sentiment analysis,” *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 120–123, 2013. DOI: <http://dx.doi.org/10.1109/ASONAM.2012.30>. [Online]. Available: <https://dl.acm.org/doi/pdf/10.5555/2456719.2456923>.
- [6] C. Comerton-Forde and T. Putnins, “Stock price manipulation: Prevalence and determinants,” *Review of Finance*, vol. 18, no. 1, pp. 23–66, 2014. [Online]. Available: <https://EconPapers.repec.org/RePEc:oup:revfin:v:18:y:2014:i:1:p:23-66..>
- [7] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, 2014.
- [8] F. Allen and D. Gale, “Stock-Price Manipulation,” *The Review of Financial Studies*, vol. 5, no. 3, pp. 503–529, May 2015, ISSN: 0893-9454. DOI: 10.1093/rfs/5.3.503. eprint: <https://academic.oup.com/rfs/article-pdf/5/3/503/24417344/050503.pdf>. [Online]. Available: <https://doi.org/10.1093/rfs/5.3.503>.
- [9] J. Danbolt, A. Siganos, and E. Vagenas-Nanos, “Investor sentiment and bidder announcement abnormal returns,” *Journal of Corporate Finance*, vol. 11, Jun. 2015. DOI: 10.1016/j.jcorpfin.2015.06.003.

- [10] K. Golmohammadi, O. Zaïane, and D. Diaz, “Detecting stock market manipulation using supervised learning algorithms,” *DSAA 2014 - Proceedings of the 2014 IEEE International Conference on Data Science and Advanced Analytics*, pp. 435–441, Mar. 2015. DOI: 10.1109/DSAA.2014.7058109.
- [11] L. M. Dang and D. Duong, “Improvement methods for stock market prediction using financial news articles,” Sep. 2016, pp. 125–129. DOI: 10.1109/NICS.2016.7725636.
- [12] R. Joshi and R. Tekchandani, “Comparative analysis of twitter data using supervised classifiers,” *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 3, pp. 1–6, 2016.
- [13] B. Yuan, “Sentiment analysis of twitter data,” 2016.
- [14] M. A. Ghazanfar, S. A. Alahmari, Y. F. Aldhafiri, A. Mustaqeem, M. Maqsood, and M. A. Azam, “Using machine learning classifiers to predict stock exchange index,” *International Journal of Machine Learning and Computing*, vol. 7, no. 2, pp. 24–29, 2017.
- [15] K. Khare, P. Gupta, V. Z. Attar, and O. Darekar, “Short term stock price prediction using deep learning,” in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology RTEICT*, IEEE; 2017, pp. 482–486, ISBN: 978-1-5090-3704-9. DOI: 10.1109/RTEICT.2017.8256643.
- [16] A. Li, J. Wu, and Z. Liu, “Market manipulation detection based on classification methods,” *Procedia Computer Science*, vol. 122, pp. 788–795, 2017, 5th International Conference on Information Technology and Quantitative Management, ITQM 2017, ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2017.11.438>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050917326868>.
- [17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” 2017. arXiv: 1706.03762 [cs.CL].
- [18] S. Loria, “Textblob documentation,” *Release 0.15*, vol. 2, 2018.
- [19] D. L. Minh, A. Sadeghi-Niaraki, H. D. Huy, K. Min, and H. Moon, “Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network,” *IEEE Access*, vol. 6, pp. 55 392–55 404, 2018.
- [20] T. Nisar and M. Yeung, “Twitter as a tool for forecasting stock market movements: A short-window event study,” *The Journal of Finance and Data Science*, vol. 4, Feb. 2018. DOI: 10.1016/j.jfds.2017.11.002.
- [21] K. Raza, “Chapter 8 - improving the prediction accuracy of heart disease with ensemble learning and majority voting rule,” in *U-Healthcare Monitoring Systems*, ser. Advances in Ubiquitous Sensing Applications for Healthcare, N. Dey, A. S. Ashour, S. J. Fong, and S. Borra, Eds., Academic Press, 2019, pp. 179–196. DOI: <https://doi.org/10.1016/B978-0-12-815370-3.00008-6>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780128153703000086>.

- [22] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news," *Journal of Ambient Intelligence and Humanized Computing*, Mar. 2020, ISSN: 1868-5145. DOI: 10.1007/s12652-020-01839-w. [Online]. Available: <https://doi.org/10.1007/s12652-020-01839-w>.
- [23] W. Khan, U. Malik, M. A. Ghazanfar, M. A. Azam, K. Alyoubi, and A. Alfakeeh, "Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis," *Soft Computing*, vol. 24, Aug. 2020. DOI: 10.1007/s00500-019-04347-y.
- [24] T. J. Putniņš, "An overview of market manipulation," *Corruption and Fraud in Financial Markets: Malpractice, Misconduct and Manipulation*, vol. 13, 2020.
- [25] Z. Zou and Z. Qu, "Using lstm in stock prediction and quantitative trading," 2020.
- [26] M. Al Guindy, "Corporate twitter use and cost of equity capital," *Journal of Corporate Finance*, vol. 68, p. 101 926, 2021, ISSN: 0929-1199. DOI: <https://doi.org/10.1016/j.jcorpfin.2021.101926>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S092911992100047X>.
- [27] C. Long, B. Lucey, and L. Yarovaya, "'i just like the stock" versus "fear and loathing on main street" : The role of reddit sentiment in the gamestop short squeeze," *SSRN Electronic Journal*, Jan. 2021. DOI: 10.2139/ssrn.3822315.