

# Stock Market Price Movement Prediction using RNN and Point-weight Sentiment

by

S. M. Rageeb Noor Uddin

17201110

Jannatul Arafat Naim

17201119

Mashuk Arefin Pranjol

17201094

Almas Ashrafi

17201111

Ibthasham Amin Emon

17201135

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

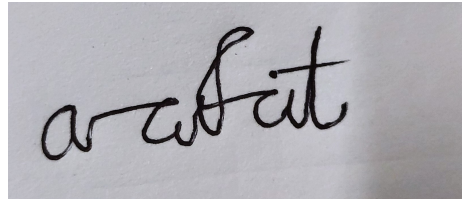
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material 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:**



---

S. M. Rageeb Noor Uddin  
17201110



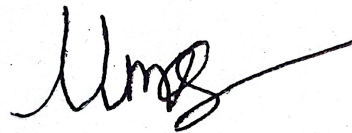
---

Jannatul Arafat Naim  
17201119



---

Mashuk Arefin Pranjol  
17201094



---

Almas Ashrafi  
17201111



---

Ibthasham Amin Emon  
17201135

# Approval

The thesis/project titled “Stock Market Price Movement Prediction using RNN and Point-weight Sentiment” submitted by

1. S. M. Rageeb Noor Uddin (17201110)
2. Jannatul Arafat Naim (17201119)
3. Mashuk Arefin Pranjol (17201094)
4. Almas Ashrafi (17201111)
5. Ibthasham Amin Emon (17201135)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 16, 2022.

## Examining Committee:

Supervisor:  
(Member)



---

Warida Rashid

Lecturer  
Department of Computer Science and Engineering  
Brac University

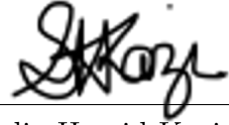
Program Coordinator:  
(Member)

---

Dr. Md. Golam Rabiul Alam

Associate Professor  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)



---

Sadia Hamid Kazi

Chairperson and Associate Professor  
Department of Computer Science and Engineering  
Brac University

# Abstract

Predicting the price of stocks has always been an exciting and challenging field for academics and investors for a long time as it helps to gain high-profit margins for investment companies, investors, and emerging advanced automated trading bots. Existing forecasting algorithms and studies on statistical models using sentiment analysis have shown promising results. However, due to the highly volatile nature of the stock market and many private and public variables that directly affect the market, it is very challenging to predict prices for extreme situations with reasonable accuracy. This study introduces a point-weight algorithm for tweets and news to gain a similar pattern as stock prices, combined with stock data and feed into the RNN network for time-series prediction. We will experiment with different mechanisms for point-weight algorithms to compare results, correlate with stock price patterns and changes while focusing on accuracy. Furthermore, we will experiment with other multivariate stocks and different architecture of RNN to find how it affects the accuracy of model training.

**Keywords:** Social media; Twitter; Newspaper; Stock Market; Prediction; Point-weight; Sentiment; RNN; LSTM; NLP; VADER;

## **Acknowledgement**

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

Secondly, we wish to express our sincerest gratitude to our advisor Warida Rashid mam for her constant support and advice in our work. She guided and mentored us relentlessly throughout our whole research.

And finally thanks to our parents and all the faculty members from BRAC University. With their kind support and prayer we are now on the verge of our graduation.

# Table of Contents

Declaration	i
Approval	ii
Abstract	iv
Acknowledgment	v
Table of Contents	vi
List of Figures	viii
List of Tables	ix
Nomenclature	x
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Research Problem . . . . .	2
1.3 Research Objectives . . . . .	4
<b>2 Literature Review</b>	<b>5</b>
2.1 Sentiment Analysis . . . . .	5
2.2 RNN . . . . .	6
2.3 LSTM . . . . .	6
2.4 Related Works . . . . .	6
<b>3 Work Plan</b>	<b>8</b>
<b>4 Datasets</b>	<b>10</b>
4.1 Input data . . . . .	10
4.2 Data Preprocessing . . . . .	12
4.3 LST) . . . . .	16
<b>5 Input Data and RNN Implement</b>	<b>18</b>
5.1 Input Data Pre-processing . . . . .	18
5.2 RNN Implementation . . . . .	18
<b>6 Result</b>	<b>20</b>
6.1 Result . . . . .	20

<b>7 Conclusion</b>	<b>23</b>
<b>bibliography</b>	<b>24</b>



# List of Figures

3.1	Work-Plan Chart . . . . .	8
3.2	Point-weight Algorithm . . . . .	9
4.1	US equities news dataset . . . . .	10
4.2	Apple historical stock data . . . . .	11
4.3	Company tweet . . . . .	11
4.4	Tweet data . . . . .	11
4.5	Joined Table . . . . .	12
4.6	Dictionary example . . . . .	13
4.7	Point data . . . . .	14
4.8	Visualization of point data table . . . . .	14
4.9	Combined data . . . . .	15
4.10	Graphical representation of combined data . . . . .	15
4.11	LSTM . . . . .	16
5.1	Model Architecture . . . . .	19
6.1	Using six features(SM Point included)) . . . . .	20
6.2	Only one feature(Closed price) . . . . .	21
6.3	Using five features(SM Point excluded) . . . . .	21

# List of Tables

6.1	Table 1 . . . . .	22
-----	-------------------	----

# Nomenclature

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

*AAPL* Code for Apple in NASDAQ

*CNN* Convolutional neural network

*GRU* Gated Recurrent Units

*LSTM* Long short-term memory

*NASDAQ* A dealer market

*NLP* Natural language processing

*NYSE* An auction market

*RNN* Recurrent neural network

*SGD* Stochastic gradient descent

*SM* Social media

*VADER* Valence Aware Dictionary for Sentiment Reasoning

# Chapter 1

## Introduction

### 1.1 Background

According to [1], Stock market or stock exchange where stock brokers and traders purchase and sell stocks, bonds, equity, or other securities. As of 2016, there are 60 stock exchanges around the world. In the year 2020, these stock exchanges had a combined market capitalisation of approximately US\$93.7 trillion. The behaviour of stock price movements are primarily affected by external factors such as socio-economic issues, inflation, exchange rate fluctuations. Other studies have found that people are preconditioned to seeing patterns and thinking in groups. Also, irrational behaviours can sway the market due to press releases, speculations, euphoria, and mass hysteria.

Social media has a considerable influence on stock prices. As devices and internet access get more affordable, social media's impact on people's lives and choices gets bigger. According to [2], Social media is used by 3.96 billion people worldwide. People usually access social media through web-based apps on their devices like laptops, desktops, and mobile devices. The majority of adults and around 86% of the younger generation use social media regularly. The most used social media apps are Facebook, Youtube, Instagram, Twitter, Tiktok, and Reddit to connect with friends, family, and others.

Twitter is one of the most popular social media platforms for sharing business-related information and opinions. According to [3], Twitter is a social network based in the United States that allows users to send and receive tweets. Twitter can be accessed through its website or its mobile device application software. Even though only 10% of its total users tweet regularly, the tweets from verified profiles are read by many people, and people trust the information shared by the verified users a lot. So, by analysing the Twitter trend, we can get information about what people think about specific topics or stocks.

News media has been the go-to place for information about current events for many years. According to [4], the news is distributed through several different methods like hearsay, publishing, postal service, radio, electronic communication, or through the testimonies of spectators and witnesses to events. Topics for news reports include warfare, government, politics, education, healthcare, the environment, economy,

commerce, fashion, entertainment, and athletic events, as well as strange or unusual occurrences. People still get information about businesses and companies through newspapers. However, the news reported on the internet, otherwise known as online news, is delivered quickly, and people find out about recent events almost in real-time. Online news makes the stock market even more volatile, and as a result, we have to keep a close eye on news media, especially online news, to increase the accuracy of stock prediction.

Twitter is mainly used for text-based updates, called "tweets". According to [5], Most celebrities and leaders use Twitter regularly, and we have seen the drastic impact of Twitter on protests or collective sentiment about any topic. Moreover, a single investor can monitor Twitter for major announcements instead of various web pages as, at present, all big companies or agencies have Twitter accounts. In addition, so many professional traders look up to Twitter to get some information about stocks.

According to [6], news on the individual firm level or national economy level significantly influences the investors. Out of the linear relationship between good news and bad news and their consequent behaviour, sometimes bad news also sheds optimistic hope for some stocks. For example, a cyclone notifying news may trigger better investment in the home insurance stocks. Furthermore, bad news for some stocks is good news for others. Professional traders utilise the news from various sources to have the upper hand in the stock market.

When interest rates are lowered, the demand for both funds and shares rises, vice versa for increased interest rates. According to [7], This is very significant for economies like the US, where the stock market has the most significant impact on the national and international economies. Changes in top management are also a notable factor in the movement of stock prices. If the investors see the change of direction positively, they perceive that the company will do better in the coming times and invest more. Drastic changes in the management trigger rumours among the investors, which often leads to short-selling or panic encashment.

According to [3], Tweets are the main form of communication on Twitter. People write and post about the topic they want to talk about, and most of the time, people use Hashtags that link the tweet to all the other posts with that hashtag [14]. This helps people from all over to join in on the conversation and give their opinions. As a result, it sets a positive or negative trend on a particular topic. Trending hashtags attract much attention from people and direct users' views in a certain way.

## 1.2 Research Problem

In this research [10], we have found the prediction report on the forecasting movement of the stock market by classifying sentiment from tweets and getting marginal value from classification algorithms. SVM performs better on classifying and predicting stock price, but high rises and falls of price have a very high error rate.

LSTM is a improved version of existing RNN, which can capture the long term

dependency in time series. According to [15], the proposed model of LSTM RNN can have a more significant impact as the model doesn't expertise much and is a time-saving model for excluding some manual steps. Alongside this, much better accuracy compared to traditional models with a better extent of data patterns. Still, there are some disadvantages, like slower forecasting speed and more resource costs than conventional models. LSTM architecture is based on artificial recurrent neural networks (RNN). It is a deep learning architecture, which can capture the long-term dependency in time series. According to [15], the proposed model of LSTM RNN can have a more significant impact as the model doesn't expertise much and is a time-saving model for excluding some manual steps. Alongside this, much better accuracy compared to traditional models with a better extent of data patterns. Still, there are some disadvantages, like slower forecasting speed and more resource costs than conventional models.

In the research [16], Deep neural network designs can capture hidden dynamics and make predictions. If a business's data is used to train the model, it will be capable of predicting that company's stock price. We also discovered that the suggested system, CNN, can detect some data interrelations. Furthermore, the results show that CNN architecture can detect changes in trends. CNN has been chosen as the best model for the suggested system. It makes predictions based on the information present at the moment. Despite the fact that RNN and LSTM are used in various time-dependent data studies, they do not outperform the CNN architecture in this case. This is related to the erratic nature of stock market fluctuations. CNN architecture outperforms RNN and LSTM in predicting the turbulent stock market.

According to [17], The research of the predictive potential of online web data is still in its early stages. The numerous connections as well as restrictions of these various data sources, as well as particular sentiment measurements and their broad application to various domains, are unclear. The comparison of different types of data is the first attempt to extract a range of sentiment indicators from several popular data sources (Twitter, search engines, and media) and use various sentiment indicators to predict different financial metrics (DJIA, trading volumes, VIX, and gold) on a weekly and daily scale using sentiment analysis. Additional research is required to understand better underlying reasons and how mood indicators are linked to and predict social and economic phenomena like stock markets.

According to [18], Stock market fluctuation prediction has traditionally been viewed as a challenging endeavour. Some technical indicators and a sentiment influence feature were utilised in a study to predict stock market fluctuation using a two-layer RNN-GRU model. The experiments suggest that the approach and attributes may accurately predict with minor mistakes. Understanding the significance of several technical analyses and selecting much more effective methods is crucial. In addition, additional text material from online social networks should be collected, and sentiment should be analysed using more advanced machine learning algorithms such as interdependent Latent Dirichlet Allocation (ILDA).

So, many studies have been done with RNN and Sentiment Analysis which produced good results. Still, we want to introduce a new algorithm, Point-weight sentiment,

and study patterns from social media to find the correlation with stock price patterns and how it affects prediction by leveraging patterns from social media.

The question this research is trying to answer is :

*How can we leverage the power of data collected from social media for Stock Market Price Prediction with the help of Deep Neural Network?*

As discussed earlier, CNN, RNN and LSTM are three main deep learning models used to train a machine to predict market prices using stock market data. We are using RNN and LSTM and implementing a new algorithm to predict the stock prices better. We intend to focus our data collection on social media, news and stock market data and combine them to predict stock prices

### **1.3 Research Objectives**

Our research aims to develop a point-weight algorithm from sentiment analysis and RNN. We will assign positive and negative sentiment points based on the positivity or negativity of the news and the number of followers of the source. We will use this to train a RNN module to predict stock prices. The objective of the research:

1. Developing a different Point-weight Algorithm mechanism from sentiment analysis.
2. Compare Stock Price with Point-weight data using mathematics and graphs
3. Feed and train RNN with point-weight results and stock data.
4. Experiment with different technical indicators to find how they affect data prediction.

# Chapter 2

## Literature Review

The stock market is indeed the backbone of a country's economy and the global economy in the capitalist world. Companies listed in the stock index raise money from general people, traders, and investors to make better products and pay workers, which is significant for growing economic and socio-conditions. Along with it, professionals like investors and traders can gain good profit from the stock market to continue the financial flow throughout every class of people in our society.

To gain profit, traders and investors require an idea of future patterns of companies performance to plan their investing. This behaviour of investors impacts companies to grow hence growing the economic condition. On the other hand, by seeing the movement of companies' stock prices, we can see behavioural patterns of a large number of people. Thus, it is crucial to gain an understanding of future patterns that helps government, industries, and investors plan more efficiently with this result and earn profits.

### 2.1 Sentiment Analysis

Sentiment Analysis is a study procedure for the subjective information about emotional state and personal info. It is widely used technique for analysing the consumer's voice in materials like reviews and survey responses, as well as online and social media and healthcare resources [9]. Feelings, attitudes, emotions, and views all fall under the heading of sentiment. Social media sentiment analysis is an excellent source of data that may help you figure out marketing tactics, increase campaign success, and so much more. In recent years, there has been a significant study on sentiment analysis on movie reviews, and Twitter feeds [10]. Sentiment analysis has been treated as a Natural Language Processing issue at different levels of abstraction. After originating as a document-level categorization assignment, it has been handled at the sentence level and, more recently, at the phrase level. At several layers of depth, sentiment analysis has been handled as a Natural Language Processing problem. It has been addressed at the sentence level, and more recently, at the phrase level, after beginning as a document level classification assignment[11].



## 2.2 RNN

RNN is an artificial neural network in which nodes are connected in a directed graph that follows a temporal sequence. This enables it to behave in a temporally dynamic manner. RNNs, is a type of feedforward neural networks, can handle variable-length sequences of inputs with the help of their hidden state (memory). As a result, tasks like unsegmented, linked handwriting recognition or speech recognition are achievable. The name "recurrent neural network" is generally used for two broad kinds of networks with a similar overall structure, one with limited impulse and the other with infinite inspiration. The action of these two types of networks is temporal and dynamic. An endless impulse recurrent network is a directed cyclic graph that cannot be unrolled and replaced with a strictly feedforward neural network. In contrast, a finite impulse recurrent network is a directed acyclic graph that can be unrolled and replaced with a strict feedforward neural network [12]. Recurrent neural networks are dynamic systems with an internal state at each classification time step. Circular connections between upper and lower layer neurons and optional self-feedback connections contribute to this. RNNs can carry data from past step to present processing step. Thus, Recollection of time series occurrences is created by RNN occurrences in this way. [13]

## 2.3 LSTM

LSTM is a deep learning architecture that implements the concept of RNN. LSTM features feedback connections, unlike conventional feedforward neural networks. It can process sequences. For example, activities like unsegmented, connected handwriting identification, speech recognition, and anomaly detection in network traffic or IDSs(intrusion detection systems) can all use LSTM[15]. RNNs are the forerunners of LSTM neural networks (RNN). RNN, a descendant of Multi-layer Perception, excels in settling sequential data. For long-term dependency concerns, RNN's performance is insufficient. The goal of LSTM is to make RNN better. LSTM has a more sophisticated inner structure than RNN. [16]

## 2.4 Related Works

This part aims to critically review previous relevant works in the field of financial market prediction, especially the works that used sentiment analysis and Recurrent Neural Networks. We have analysed different techniques used in studies that have been used to predict future prices of companies, thus understanding behavioural patterns.

In this study [10], they have labelled Twitter data to predict classification where they have found, SVM performed best with Unigram parameters. After that, they have taken the average marginal value from SVM for all the tweets from that day to predict stock prices. Though they expected stock prices, their main focus was sentiment analysis in this study.

The research work [17] experimented with the multi-step time series forecasting

problem and proposed several modelling solutions based on various data patterns. They also conducted tests to evaluate the efficacy of their suggested strategy, which included LSTM RNN for multi-step processes and compared the findings to those obtained using the ARIMA/GRNN model. Although their multi-step forward prediction findings were not highly accurate, they did have a reference value.

The research work [18] proposed a formalisation for stock price prediction based on deep learning. Their proposed method is a model-independent approach. Instead of fitting data to a specific model, they identified deep learning architectures used to uncover latent dynamics in data. For the price prediction of NSE listed businesses, they employed three distinct deep learning architectures (LSTM, RNN, and CNN).

The paper [19] surveyed a wide range of online data sets and ways for tracking sentiment to predict the financial market. They compared their predictions with market indices. DIJA, VIX, and volume prediction discovered increases in direction accuracy and MAPE of prediction performance. The improvement, however, was not significant.

In the research paper [20], To anticipate stock volatility in the Chinese stock market, they developed a model based on RNN with GRU. They suggested specific technical indicators and a sentiment influence feature, which they then used to forecast stock volatility using a two-layer RNN-GRU model. They discovered that their model could accurately predict with minor mistakes.

The research [21] tried to predict how volatile Bitcoin's price is By studying the sentiments on Twitter. They collected tweets about bitcoin and classified them into positive and negative sentiments. The feelings were then fed into an RNN model and historical Bitcoin prices to forecast future Bitcoin prices. Their overall accuracy of the price was 77.62%.

# Chapter 3

## Work Plan

The purpose of using Recurrent Neural Network (RNN) and the point-weight sentiment is to predict stock market price movement more accurately and find the correlation. In order to do so, the algorithm requires data of previous stock market prices of the organisation, news and social chatter about the organisation. Then, the algorithm will predict the organisation's future stock market price. Figure 3.1 provides an overview of the processes for this study.

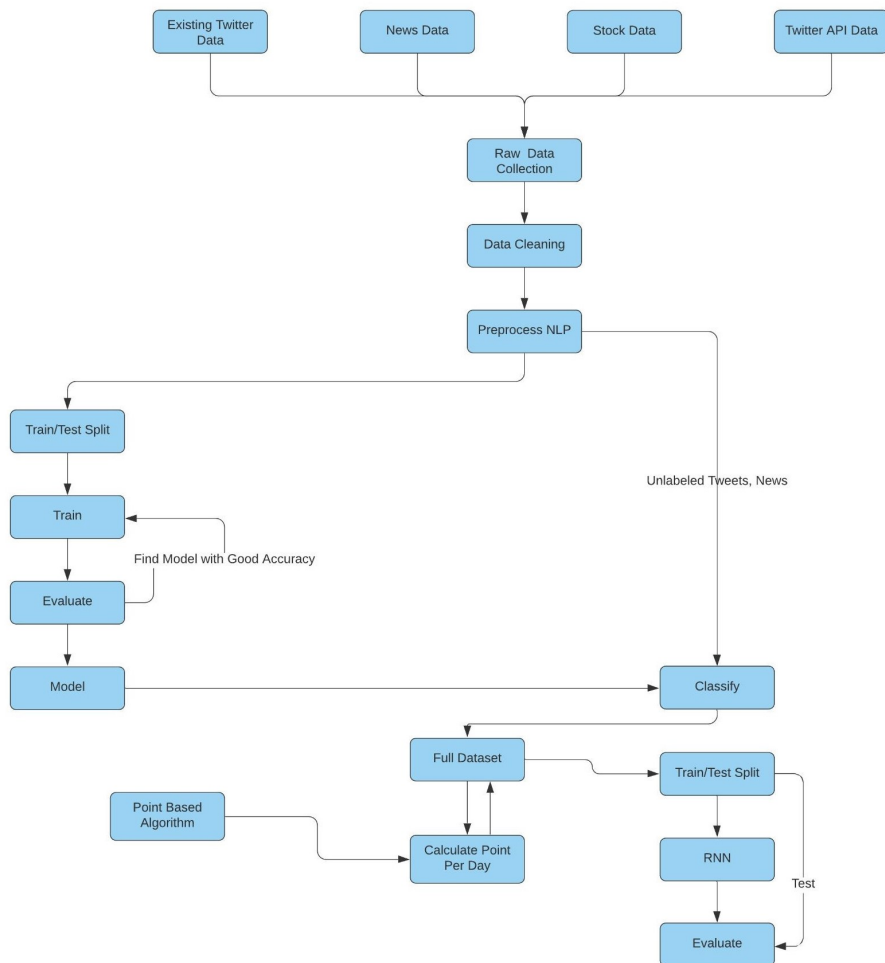


Figure 3.1: Work-Plan Chart

Existing Tweets, News, Stocks, Twitter API data goes into the raw data collection. After data cleaning and preprocessing(NLP), existing Twitter data gets split into train and test splits. The train split gets evaluated with a model with reasonable accuracy, and that model data and unlabeled tweets, the news gets classified and goes into the entire dataset. Moreover, after calculating points per day, the data from the point-weight algorithm goes into the whole dataset. The proposed mechanism for a Point-weight Algorithm can be found in Figure 3.2,

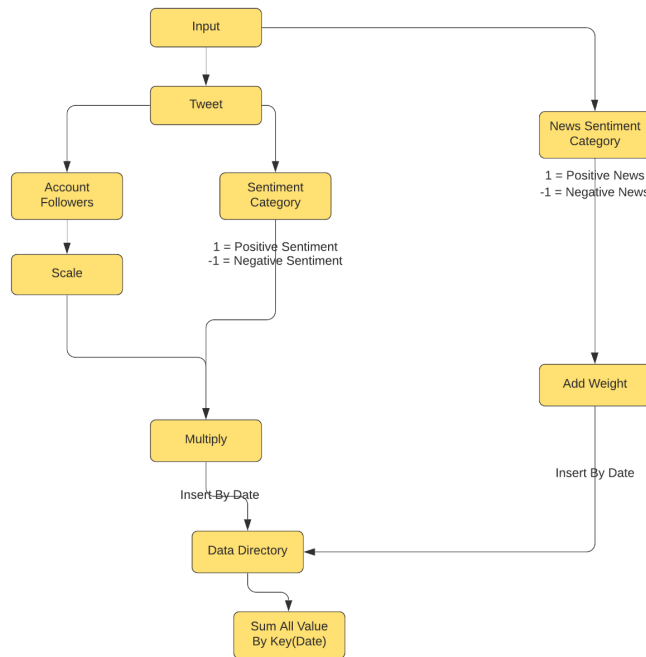


Figure 3.2: Point-weight Algorithm

For the point-weight algorithm, the tweets get scaled by account followers. Also, positive tweets get assigned as "1" sentiment, and negative tweets get "-1" assigned. After multiplying the sentiments with the scales, the tweets will be inserted into the data dictionary by date. The news also works in the same way that positive news gets assigned "1" and negative news gets assigned with "-1". Then fixed weights differ by news source popularity are added to the news. Then, it will be inserted into the data dictionary by date. Then, all values will be summed up by key from the dictionary(date).

The data from the entire dataset gets split into train and test splits, and then the data goes into the RNN algorithm to train with multivariate and gets evaluated with test data.

# Chapter 4

## Datasets

### 4.1 Input data

For input data we used three datasets. As the first dataset we used a dataset we found on “kaggle.com” named Historical financial news archive[22]. This dataset contains news archives of more than 800 U.S companies from 2008 to 2020. This data represents the historical news archives of the US equities publicly traded on NYSE/NASDAQ that have above 10\$ price for each share. After filtering only the data of Apple company the total data we used is 20231. We filtered by ticker(AAPL) and only used content and release date from this dataset. In Figure 4.1 we can see an example of this dataset,

Unnamed: 0	id	ticker	title	category	content	release_date	provider	url	article_id	
0	49183	270698	AAPL	JPMorgan cautious ahead of Apple earnings	news	JPMorgan lifts its Apple AAPL 2.9 target f...	2020-01-28	Seeking Alpha	https://invst.ly/pnjv8	2068762
1	49184	270699	AAPL	FAANG s Fall but Get Some Wall Street Love	news	By Kim Khan \nInvesting.com The FAANG stocks...	2020-01-28	Investing.com	https://www.investing.com/news/stock-market-ne...	2068765
2	49185	270700	AAPL	Wall Street tumbles as virus fuels economic worry	news	By Chuck Mikolajczak NEW YORK Reuters U S ...	2020-01-28	Reuters	https://www.investing.com/news/stock-market-ne...	2068311
3	49186	270701	AAPL	Earnings Watch Apple and AMD to take earnings...	news	Two of the best performing tech stocks of 2019...	2020-01-28	MarketWatch	https://invst.ly/pnlbs	2068906
4	49187	270702	AAPL	Day Ahead Top 3 Things to Watch for Jan 28	news	By Yasin Ebrahim and Kim Khan \n1 Apple Readi...	2020-01-28	Investing.com	https://www.investing.com/news/stock-market-ne...	2068907

Figure 4.1: US equities news dataset

Then we used Apple (AAPL) Historical Stock Data. Another dataset we found on “kaggle.com”[23]. This dataset contains Apple’s (AAPL) stock data from 2010 to 2020. It contains all daily statistics related to Apple’s(AAPL) trading since 2010 including closing price, opening price, maximum price and volume. The number of total data we used from here is 2323. From this dataset we only took the data of news and date. This date is shown in Figure 4.2.

	Date	Open	High	Low	Close	Adj Close	Volume	News
0	2006-12-01	13.114285	13.190000	12.871428	91.320000	13.045714	198769900	WHAT'S ON TONIGHT : 8 P.M. (TLC) ASHLEY JUDD A...
1	2006-12-04	13.125714	13.150000	12.928572	91.120003	13.017143	177384200	More on Housing Prices : The broadest governme...
3	2006-12-06	12.948571	13.055715	12.810000	89.830002	12.832857	159546100	Honoring R.W. Apple in Words and Food : About ...
4	2006-12-07	12.861428	12.928572	12.414286	87.040001	12.434286	251206900	Homebuilders, and Worries Over Jobs, Lead a De...
5	2006-12-08	12.461429	12.770000	12.428572	88.259995	12.608571	196069300	Homebuilders, and Worries Over Jobs, Lead a De...

Figure 4.2: Apple historical stock data

The final dataset we used is named Tweets about the top companies from 2015 to 2020 [24]. This dataset, as a part of the paper published in the 2020 IEEE International Conference on Big Data under the 6th Special Session on Intelligent Data Mining track, is created to determine possible speculators and influencers in a stock market. This dataset contains over 3 million unique tweets with their information such as tweet id, author of the tweet, post date, the text body of the tweet and the number of comments, likes, and retweets of tweets matched with the related company. From this dataset we filtered data only for Apple. In Figure 4.3 and Figure 4.4 the company tweet and tweet data is shown,

	tweet_id	ticker_symbol
0	550803612197457920	AAPL
1	550803610825928706	AAPL
2	550803225113157632	AAPL
3	550802957370159104	AAPL
4	550802855129382912	AAPL

Figure 4.3: Company tweet

	tweet_id	writer	post_date	body	comment_num	retweet_num	like_num
0	550441509175443456	VisualStockRSRC	1420070457	lx21 made 10, 008onAAPL -Check it out! htt...	0	0	1
1	550441672312512512	KeralaGuy77	1420070496	Insanity of today weirdo massive selling. \$aap...	0	0	0
2	550441732014223360	DozenStocks	1420070510	S&P100 #Stocks Performance <i>HDLOW SBUXTGT</i> ...	0	0	0
3	550442977802207232	ShowDreamCar	1420070807	<i>GMTSLA</i> : Volkswagen Pushes 2014 Record Recal...	0	0	1
4	550443807834402816	i_Know_First	1420071005	Swing Trading: Up To 8.91% Return In 14 Days h...	0	0	1

Figure 4.4: Tweet data

Then we joined these two tables based on the column "tweet\_id". In the Figure 4.5 we can see the joined table.

	tweet_id	ticker_symbol	writer	post_date	body	comment_num	retweet_num	like_num
0	550803612197457920	AAPL	SentiQuant	2015-01-02	#TOPTICKERTWEETS AAPLIMRS BABAEBAY \$AMZN...	0	0	1
2	550803610825928706	AAPL	SentiQuant	2015-01-02	#SENTISHIFTUP KFB GOOGLGS GOLDT \$AAPL...	0	0	1
5	550803225113157632	AAPL	MacHashNews	2015-01-02	Rumor Roundup: What to expect when you're expe...	0	0	0
6	550802957370159104	AAPL	WallLightShed	2015-01-02	An \$AAPL store line in Sapporo Japan for the "...	2	4	4
7	550802855129382912	AAPL	2ways trading	2015-01-02	AAPL - Will AAPL Give Second entry opportuni...	0	0	0

Figure 4.5: Joined Table

The total number of rows we used from here is 1425013.

## 4.2 Data Preprocessing

Sentiment Analysis is the process of finding out computationally if a written article is positive, negative or neutral. Sentiment Analysis is helpful in many sectors. For example, it can be used to monitor and analyse human behaviour, detect possible dangerous situations, and determine a general media mood. For Sentiment Analysis in our research paper, we used vaderSentiment. VADER is a lexicon and rule-based sentiment analysis tool that is accommodated explicitly to sentiments shown in social media. VADER is fully open-sourced under the MIT license. By using VADER, we can tell about the positivity or negativity scale of the output. Many statements in the news and on Twitter about the stock prices, but how positive or negative the statements are can be known by using Vader. In vaderSentiment, the scoring works like a compound score computed by adding up the valence scores of each word in the lexicon, adjusted according to the rules, and then normalised to be between -1 (most extreme negative) and +1 (most extreme positive). We found that it is the most useful metric if one wants a single unidimensional measure of sentiment for a given sentence.

The threshold values we used are:

1. Positive sentiment: compound score  $>0$
2. Negative sentiment: compound score  $\leq 0$

Firstly, we get all the dates for all three datasets (news and Twitter). Then we get all the unique dates, and using those dates; we created a dictionary where the key is date. We append tweets and news in dictionary format to the key "date" for each date. This dictionary holds content, sentiment classification using vaderSentiment, weight, extra weight. An example of the dictionary is given in Figure 7 below. In the case of weights, we gave every news a weight of 5 and every tweet a weight of 1. In the case of extra weight we gave every news item an extra weight value of 1. In the case of future modifications, we will modify this value and base it on the news source credibility and popularity. We will count the value of extra weight for tweets based on the number of comments, retweets, and likes.

Here is the calculation:

$$ExtraWeight = (CommentNum + RetweetNum + LikeNum)/3$$

```
{'2021-01-01': [{'text': 'News Content',
  'sentiment': -1,
  'weight': 5,
  'extra_weight': 1},
{'text': 'Tweet Content',
  'sentiment': -1,
  'weight': 1,
  'extra_weight': 2.67}],
'2021-01-02': [{'text': 'News Content',
  'sentiment': -1,
  'weight': 5,
  'extra_weight': 1},
{'text': 'Tweet Content',
  'sentiment': -1,
  'weight': 1,
  'extra_weight': 2.67}]}
```

Figure 4.6: Dictionary example

Furthermore, we created a new table with Date and Point as columns. We calculated the points by using all the values (sentiment, weight and extra weight) for Each Date (Dictionary key). We used this equation to calculate the points for each date.

*PointCalculation* =

$$\sum (sentiment \times (weight \times extraweight))$$

The shape of the final point dataset is (1854, 2). Figure 4.7 is an example of the Point Date table.



	<b>date</b>	<b>point</b>
<b>2022</b>	2015-01-01	274.666667
<b>2023</b>	2015-01-02	-13.666667
<b>2024</b>	2015-01-03	-46.333333
<b>2025</b>	2015-01-04	124.333333
<b>2026</b>	2015-01-05	-39.000000

Figure 4.7: Point data

In the Figure 4.8, we see the graphical representation of the point data table from 2015 to 2020.

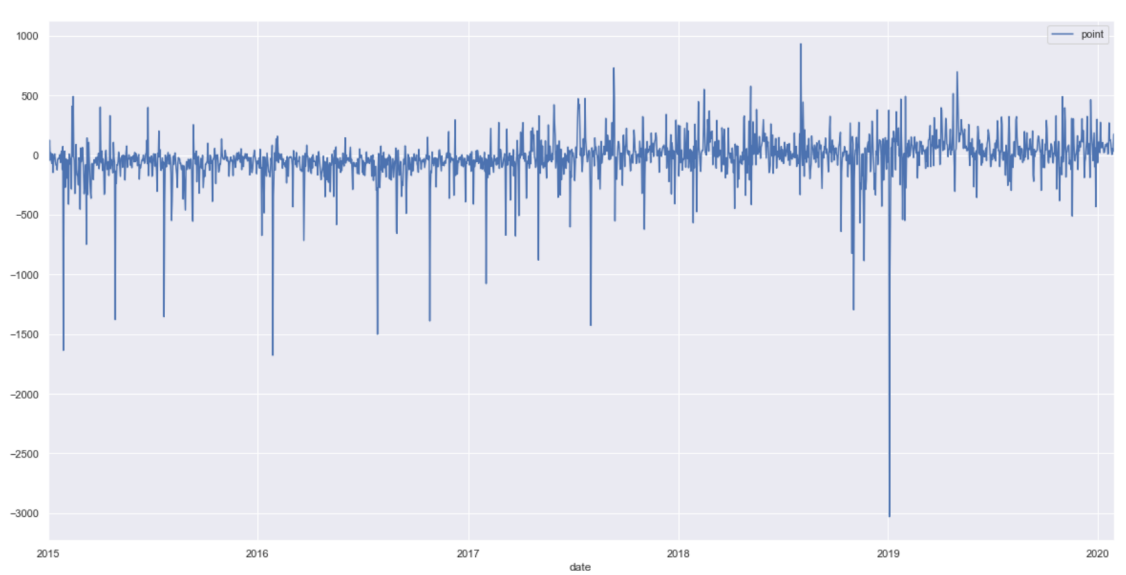


Figure 4.8: Visualization of point data table

Next, we worked with the APPLE stock data. For this table, we converted the Date into Y-M-D format. Furthermore, we removed the dollar sign to turn the prices into

proper numeric values. Next, we combined the Point Data and Stock Data based on date. An example of the combined data is shown in Figure 4.9.

	date	Close	Volume	Open	High	Low	point
1275	2015-01-02	109.33	53143770	111.39	111.44	107.3500	-13.666667
1274	2015-01-05	106.25	64210880	108.29	108.65	105.4100	-39.000000
1273	2015-01-06	106.26	65682250	106.54	107.43	104.6300	19.000000
1272	2015-01-07	107.75	39983350	107.20	108.20	106.6950	-46.333333
1271	2015-01-08	111.89	59168760	109.23	112.15	108.7000	-69.666667
...	...	...	...	...	...	...	...
4	2020-01-22	317.70	25458120	318.58	319.99	317.3100	130.000000
3	2020-01-23	319.23	26117990	317.92	319.56	315.6500	100.000000
2	2020-01-24	318.31	36634380	320.25	323.33	317.5188	40.000000
1	2020-01-27	308.95	40485010	310.06	311.77	304.8800	30.000000
0	2020-01-28	317.69	40558490	312.60	318.40	312.1900	175.000000

1276 rows × 7 columns

Figure 4.9: Combined data

Graphical representation of the Combined Data looks like the Figure 4.10.



Figure 4.10: Graphical representation of combined data

The graph shows that the stock price is correlated with the SM point (social media point). We can see that the stock price increases or decreases on many dates whenever the SM point increases or decreases. In 2019 we can see an extreme case where both stock price and SM point fell drastically.

### 4.3 LST)

RNNs are dynamic systems; each stage of the classification has an internal state. This is because of the circular links between neurons in higher and lower classes and optional self-feedback. The feedback links allow RNNs to spread data from previous events to the current processing steps. RNNs then generate a memory of events in a time series. We used RNN in this project. Sometimes, only recent information needs to be examined to carry out the present task. In theory, Long-term dependencies can be managed by the RNN. But in practice, when the range between data connections increases, RNNs are incapable of learning how to connect data. And this problem is not present in LSTMs. That's why we specifically used Long Short-Term Memory. LSTM networks are a special RNN, which can learn long-term dependencies. The long term recall of information is virtually their default behaviour and not something they struggle to understand.

A diagram of Long Short-Term Memory is shown in Figure 4.11

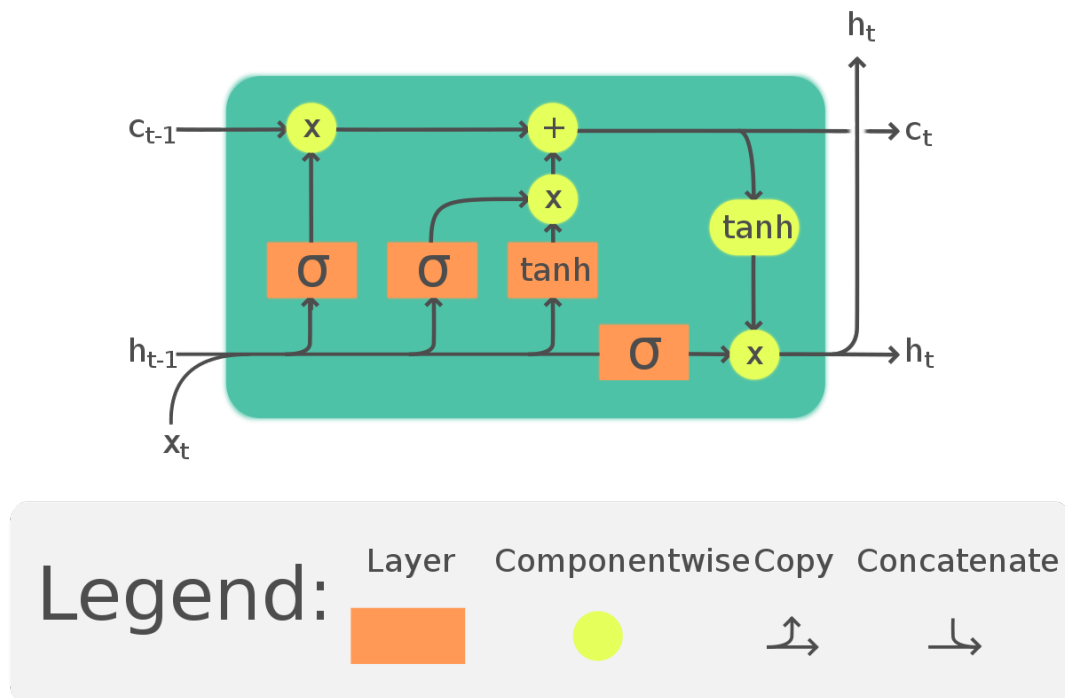


Figure 4.11: LSTM

The compact forms of the equation of LSTM are :

$$\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c}_t &= \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}$$

Here, the initial states are  $C_0 = 0, h_0 = 0$  and the operator  $\circ$  are the hadamard product. subscript t indexing time step.

The variables are:

$x_t \in R^h$  : input vector for the LSTM unit

$f_t \in (0, 1)^h$  : forget gate activation vector

$i_t \in (0, 1)^h$  : input/update gate activation vector  $o_t \in (0, 1)^h$  : output gate activation vector  $h_t \in (-1, 1)^h$  : hidden state or output vector LSTM unit

$\tilde{c}_t \in (-1, 1)^h$  : cell input activation vector

$c_t \in R^h$  : cell state vector

$\sigma_g$  : sigmoid function

$\sigma_c$  : hyperbolic tangent function

$W \in R^{h \times d}, U \in R^{h \times h}, b \in R^h$  : weight matrices and bias vector parameters that requires to be learned while training.

Here, the superscripts d and h, representing the number of input features and number of hidden units.

# Chapter 5

## Input Data and RNN Implement

### 5.1 Input Data Pre-processing

From the above mentioned Combined Dataset, we take all the columns except the Date column to make a new table and use MinMaxScaler. What MinMaxScaler does is for every value in a feature, it subtracts the minimum value in the feature and divides by the range. The range is the difference between the original maximum and the original minimum. Using MinMaxScaler we scale all of the data between 0-1. Next, we divide the data into train sets. For the training set, there will be two parts, one is X\_train, and another one is y\_train. For X\_train, for each iteration, we take 60 steps of data and for y\_train 60+1 nth number of data. For each iteration, the number of iterations is increased by 1, and it will iterate until the end of the row. The shape of the X\_train for our case is (1166, 60, 1).

### 5.2 RNN Implementation

For our research, we are using Keras as the deep learning library for implementing RNN. Keras was developed with the focus on warranting fast experimentation, which is perfect for us as we need to work with massive amounts of data. The faster we can process the data, the better the result will be. We use Keras to initialise the model architecture using the Keras Sequential class. This class provides training and inference features on this model. Next, we added four layers of LSTM, each containing 50 units. In LSTM, input shape and size will be the total number of rows and total features. For the last layer, we are getting the shape of 1 or a single value.

We are using Adam optimiser as the optimiser. Adam is a replacement optimisation algorithm for stochastic gradient descent training deep learning models. Stochastic gradient descent is an iterative method for optimising an objective function with suitable smoothness properties. Adam uses the best properties of the AdaGrad and RMSProp algorithms to provide an optimisation algorithm that can handle sparse gradients on noisy problems. Furthermore, we are using a mean squared error function to handle loss. Mean squared error (MSE) is widely used as the loss function for regression. The loss is the mean overseen data of the squared differences between actual and predicted values. For our research, we are using Total epoch 100, and the batch size is 32. The Figure 5.1 below is of a model architecture.

Layer (type)	Output Shape	Param #
lstm_5 (LSTM)	(None, 60, 50)	11400
dropout_5 (Dropout)	(None, 60, 50)	0
lstm_6 (LSTM)	(None, 60, 50)	20200
dropout_6 (Dropout)	(None, 60, 50)	0
lstm_7 (LSTM)	(None, 60, 50)	20200
dropout_7 (Dropout)	(None, 60, 50)	0
lstm_8 (LSTM)	(None, 50)	20200
dropout_8 (Dropout)	(None, 50)	0
dense_2 (Dense)	(None, 1)	51
Total params: 72,051		
Trainable params: 72,051		
Non-trainable params: 0		

Figure 5.1: Model Architecture

# Chapter 6

## Result

### 6.1 Result

We used the 6 features we used to make the combined data (including Social Media Point) to get our results. For the final epoch we have the loss value  $6.3072e-04$ . As a test set, we are using the next 50 days date where the date for the training set ends. After predicting with model and inverse transforming the result, we get results like the Figure 6.1 below,



Figure 6.1: Using six features(SM Point included))

For the next case, we have only used 1 feature which is the closed price to train and predict which we can see in Figure 6.2.



Figure 6.2: Only one feature(Closed price)

For the last case, we have used 5 features which exclude SM Point data to see how SM Point data influence the accuracy. The visualization is this result shown in Figure 6.3.



Figure 6.3: Using five features(SM Point excluded)



To calculate accuracy rate for each prediction we are using Mean Absolute Error. Mean Absolute Error is a measure of errors between paired observations expressing the same phenomenon. Equation to calculate Mean Absolute Error is,

*MeanAbsoluteError* =

$$\sum |(RealValue - PredictedValue)|/TotalInput$$

In Table 1, we can find about accuracy and final loss of train for each case.

Total Feature	Final Loss	Mean Absolute Error
6	6.3072e-04	3.50
1	6.3478e-04	6.649
5	5.3911e-04	5.90

Table 6.1: Table 1

# Chapter 7

## Conclusion

For this research, the purpose was to see the influence of social media in stock price movement rather than having the best accuracy. From the result section from Table 1, we can see that for total feature 6 (SM Point included) has the lowest Mean Absolute Error following to total feature five, then total feature 1. The lowest Mean Absolute Error means a higher accuracy rate than the other 2.

Stock price prediction using deep learning has always been one of the lucrative research points for data scientists and researchers. Although much research has focused on stock price prediction using various deep learning methods, it is still impossible to accurately predict stock prices in extreme situations. CNN, RNN and Sentiment Analysis are the main modules that are used in stock prediction, and we are going to use two of them along with a new algorithm that we are going to introduce in this research in the hopes that we can improve the accuracy of stock prediction more than what has already been done.

In future, we can optimise different areas of this process to increase the accuracy rate to the maximum point. For example, we are diversifying more social and news data, having more sentiment accuracy, and having more social data overall. We can also fine-tune our RNN model architecture.

# Bibliography

- [1] DESJARDINS, J. (2016, February 16). All of the World's Stock Exchanges by Size. <http://money.visualcapitalist.com/all-of-the-worlds-stock-exchanges-by-size/>
- [2] Robb, Michael B. (2020). Teens and the News: The Influencers, Celebrities, and Platforms They say Matter Most, 2020 <https://www.common sense media.org/research/teens-and-the-news-the-influencers-celebrities-and-platforms-they-say-matter-most-2020>
- [3] Wojcik, S., Hughes, A. (April 24, 2019). Sizing Up Twitter Users. <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>
- [4] Silverblatt Zlobin. (2004). International Communication.
- [5] Twitter, and the importance of financial tweets and news to investors and traders. (2021). Cityfalcon.Zohosites.Com. <http://cityfalcon.zohosites.com/twitter-tweets-important-for-financial-investors-traders>
- [6] BEERS, B. (2021, January 9). How the News Affects Stock Prices. Investopedia. <https://www.investopedia.com/ask/answers/155.asp>
- [7] Factors that can affect stock prices — stocks. (2018, February 6). GetSmarter-AboutMoney.Ca. <https://www.getsmarteraboutmoney.ca/invest/investment-products/stocks/factors-that-can-affect-stock-prices/>
- [8] Amadeo, K. (2020, November 27). How Stock Investing Affects the US Economy. The Balance. <https://www.thebalance.com/how-do-stocks-and-stock-investing-affect-the-u-s-economy-3306179>
- [9] Hamborg, F., Donnay, K.(2021). NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles. <https://aclanthology.org/2021.eacl-main.142/>
- [10] Chakraborty, P., Priya, U. S., Rony, M. R. A. H., Majumdar, M. A. (2017). Predicting stock movement using sentiment analysis of Twitter feed. 2017 6th International Conference on Informatics, Electronics and Vision 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT). Published. <https://doi.org/10.1109/iciev.2017.8338584>
- [11] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
- [12] Miljanovic, M. (2012, Feb-Mar). Comparative analysis of Recurrent and Finite

Impulse Response Neural Networks in Time Series Prediction. <http://www.ijcse.com/docs/INDJCSE12-03-01-028.pdf>

[13] Staudemeyer, R. C., Morris, E. R. (2019). Understanding LSTM—a tutorial into Long Short-Term Memory Recurrent Neural Networks.

[14] Fishman, E. (n.d.). How to create and use hashtags. Twitter. Retrieved May 30, 2021, from <https://business.twitter.com/en/blog/how-to-create-and-use-hashtags.html>

[15] Sak, H. , Senior, A. , Beaufays, F. (2014). Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. <https://web.archive.org/web/20180424203806/https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43905.pdf>

[16] Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., ... Cheng, Z. (2020). Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *Journal of Petroleum Science and Engineering*, 186, 106682.

[17] Yunpeng, L., Di, H., Junpeng, B., Yong, Q. (2017, November). Multi-step ahead time series forecasting for different data patterns based on LSTM recurrent neural network. In 2017 14th web information systems and applications conference (WISA) (pp. 305-310). IEEE.

[18] Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., Soman, K. P. (2017, September). Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) (pp. 1643-1647). IEEE.

[19] Mao, H., Counts, S., Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051.

[20] Chen, W., Zhang, Y., Yeo, C. K., Lau, C. T., Lee, B. S. (2017, September). Stock market prediction using neural networks through news on online social networks. In 2017 international smart cities conference (ISC2) (pp. 1-6). IEEE.

[21] Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., Lama, B. K. (2018, October). Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS) (pp. 128-132). IEEE.

[22] Historical financial news archive. (2021). Retrieved 30 September 2021, from <https://www.kaggle.com/gennadiyr/us-equities-news-data>

[23] Paparaju, T. (2020). Apple (AAPL) Historical Stock Data. Retrieved 30 September 2021, from <https://www.kaggle.com/tarunpaparaju/apple-aapl-historical-stock-data>

[24] Metin, Ö., Dogan, M. (2020). Tweets about the Top Companies from 2015 to 2020. Retrieved 30 September 2021, from <https://www.kaggle.com/omermetinn/tweets-about-the-top-companies-from-2015-to-2020?select=Tweet.csv>