A Descriptive Documentation On Stock Market Analysis Using Sentimental Analysis of Twitter Data

**ANKIT GOUR**

2982713

Submitted in partial fulfilment for the degree of

Master of Science in Big Data Management and Analytics

Griffith College Dublin

September 2020

Under the supervision of

Prof. Aqeel Kazmi

**Disclaimer**

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Applied Digital Media at Griffith College Dublin, is entirely my work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

**Signed: ANKIT GOUR Date: 06/09/2020**

# Acknowledgements

I would especially like to thank my supervisor Aqeel Kazmi for his tremendous guidance and support, which he provided throughout the research process. His wisdom and high intellect helped me a lot in grooming my project and making this project what it is now.

I want to start with a special thanks to my family members and friends who have encouraged and motivated me throughout this MSc course. Their support through the trials and tribulations over the years has made the completion of this course possible.

I would like to thank my colleague, who afforded me their time and knowledge for this dissertation. Sincere gratitude is given to all my responders whose insights and opinions helped me figure out my findings and helped me proceed in my dissertation.

Finally, I wish to acknowledge Griffith College Dublin, and its supporting team members, the industry and the resources provided by them, which taught me a lot this ear throughout the curriculum.

Table of Contents

[Acknowledgements iii](#_Toc50537597)

[List of Figures v](#_Toc50537598)

[List of Tables v](#_Toc50537599)

[List of Equation vi](#_Toc50537600)

[Abstract vii](#_Toc50537601)

[Chapter 1 . Introduction 1](#_Toc50537602)

[1.1 Your Area: 1](#_Toc50537603)

[1.2 Overview of Approach: 2](#_Toc50537604)

[1.3 Problem Statement: 3](#_Toc50537605)

[1.5 Structure of the document: 4](#_Toc50537606)

[Chapter 2 . Background 6](#_Toc50537607)

[2.1 Introduction: 6](#_Toc50537608)

[2.2 Literature Review: 6](#_Toc50537609)

[2.3 Twitter Overview: 7](#_Toc50537610)

[2.4 Related Work: 14](#_Toc50537611)

[Chapter 3 . Methodology 16](#_Toc50537612)

[3.1 Overview: 16](#_Toc50537613)

[Chapter 4 . System Design and Specifications 21](#_Toc50537614)

[4.1 Software Specification: 21](#_Toc50537615)

[4.2 Hardware Configuration: 21](#_Toc50537616)

[4.3 Libraries used were: 21](#_Toc50537617)

[Chapter 5 . Implementation 25](#_Toc50537618)

[5.1 Architecture: 25](#_Toc50537619)

[5.2 Data pre-processing: 25](#_Toc50537620)

[5.3 Data Visualization: 26](#_Toc50537621)

[5.4 CNN based Model: 26](#_Toc50537622)

[5.5 User Interface: 26](#_Toc50537623)

[Chapter 6 . Results 27](#_Toc50537624)

[6.1 Introduction: 27](#_Toc50537625)

[6.2 Results: 29](#_Toc50537626)

[Chapter 7 . Conclusion & Future Work 31](#_Toc50537627)

[7.1 Conclusion: 31](#_Toc50537628)

[7.2 Future Work: 31](#_Toc50537629)

[Chapter 8 References 32](#_Toc50537630)

# List of Figures

[Figure 1.1(CHEN, 2020) Neural Network 2](#_Toc50537633)

[Figure 2.1 **(Lei & Wang, 2015)** 11](#_Toc50537634)

[Figure 2.2(**Lei & Wang, 2015)** 11](#_Toc50537635)

[Figure 2.3(**Lei & Wang, 2015)** 12](#_Toc50537636)

[Figure 2.4(**Lei & Wang, 2015)** 13](#_Toc50537637)

[Figure 2.5 (Jang, et al., 2019) general CNN model 13](#_Toc50537638)

[Figure 2.6 (Jang, et al., 2019) 14](#_Toc50537639)

[Figure 3.1 (BRITZ, 2015) matrix and filter 16](#_Toc50537640)

[Figure 3.2 (BRITZ, 2015) CNN architecture for NLP 17](#_Toc50537641)

[Figure 3.3(BRITZ, 2015) Narrow and wide convolution 18](#_Toc50537642)

[Figure 3.4 (BRITZ, 2015) 19](#_Toc50537643)

[Figure 3.5 (BRITZ, 2015) 19](#_Toc50537644)

[Figure 4.1Training Phase flow diagram 23](#_Toc50537645)

[Figure 4.2Testing flow diagram 23](#_Toc50537646)

[Figure 5.1 Architecture 25](#_Toc50537647)

[Figure 5.2 UI design 26](#_Toc50537648)

[Figure 6.1 Prediction result 29](#_Toc50537649)

[Figure 6.2 Sentiment result 29](#_Toc50537650)

# List of Tables

[Table 2.1 (**Yin, et al., 2017)** Performance report of CNN and other KNN algorithms 10](#_Toc50537651)

[Table 4.1Software specification 21](#_Toc50537652)

[Table 4.2 Hardware Configuration 21](#_Toc50537653)

# List of Equation

[Equation 2.1 11](#_Toc50537654)

[Equation 2.2 12](#_Toc50537655)

[Equation 3.1 18](#_Toc50537656)

[Equation 3.2 18](#_Toc50537657)

[Equation 3.3 18](#_Toc50537658)

[Equation 6.1 27](#_Toc50537659)

[Equation 6.2 27](#_Toc50537660)

[Equation 6.3 28](#_Toc50537661)

[Equation 6.4 28](#_Toc50537662)

[Equation 6.5 28](#_Toc50537663)

[Equation 6.6 28](#_Toc50537664)

[Equation 6.7 28](#_Toc50537665)

[Equation 6.8 29](#_Toc50537666)

[Equation 6.9 30](#_Toc50537667)

# Abstract

The stock market has been around right from the 16th century. When east India Company first applied the concept of a Stock market for funding their voyages. The stock market has been a great attraction for people who have some money, to begin with, and want a quick profit out of it. But investing in new business or already running companies can be a double-edged sword. A loss in business will result in the loss of investors’ money. It’s not easy to predict the value of the stock’s future. The stock market can have its fair share of ups and downs with a highly unreliable pattern of stock values. The Stock market is indirectly dependent on the view of the public about the company. The tweeter is one of the leading public view platforms which give knowledge and bring the outcome how the people are reacting on the company as there is so many news about the organization and the conversation of public on the company. This thesis aims at developing a supervised machine learning model for predicting future stock prices based on public views using twitter sentiment analysis. A microblogging social platform like twitter can help us to gain the mass understanding of the values of the stocks which our model will make use for predicting future stock prices.

# . Introduction

## 1.1 Your Area:

After the Internet Revolution, social media sites have come to its existence. The social media sites have acquired an important place in almost every human life. Spending time on social media and interacting with them has become a habit for every human being. Social media has bought the world together. Anything happening in one corner of the world gets known to people sitting on the other corner. News spread faster as well as the information. Twitter consists of almost 100 million active users and about 500 million tweets every day. Twitter is a platform where general masses can share their opinion on a topic. In recent years, stock market-related information and people's views are also being discussed on twitter. EMH (Efficient Market Hypothesis) from their research states that information is a crucial factor behind the stock market prices. Several researchers have attempted to find any pattern to predict the stock prices by using its behaviour and its response to external stimuli. In 1929 the first-ever most tremendous stock market crash occurred also known as Black Tuesday. It lasted for about one month from September to October. The first-ever devastating stock market crash ever happened. The recent stock market crash occurred in 2020 from 24th February to 28th February. Regardless of stock market crashes, the price of stocks continuously deviates.

* **Neural network** (CHEN, 2020)**:**

Neural networks are a set of models generated which tries to find out the existing underlying patterns in a dataset which can help in making the final prediction. Neural network algorithms tend to mimic human brain functions to reach a particular conclusion. It tries to execute operations similarly a human brain neuron would do. Hence, the Neural name network. Neural network-based algorithms consist of a system of neurons. It has the ability to adapt to the changing inputs. Therefore, the network is known to produce the best results in any condition.

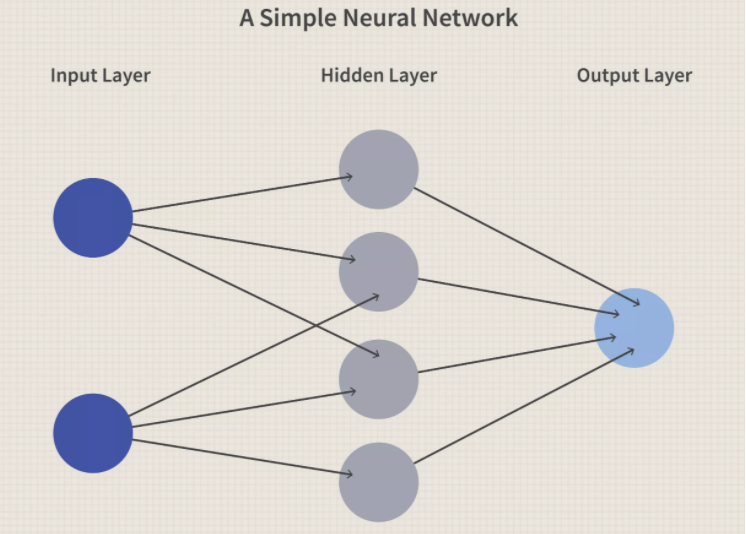


Figure .(CHEN, 2020) Neural Network

**KEY TAKEAWAYS**

* Neural network algorithms in machine learning are completely based on the function of actual human-based neurons.
* To its ability to deal with continuously changing dataset, it is used in a variety of sectors like financial services, forecasting, marketing, fraud detection, risk management etc.
* Neural networks are generally going to algorithms for predicting stock market prices.

## 1.2 Overview of Approach:

This thesis will make use of twitter tweets to obtain information about a company stock by various views of the public. Concepts of Natural language processing will be used for extracting sentiments from the tweets. A machine learning model will be build using supervised machine learning algorithms, and predictions will be made about future stock market prices.

As we have discussed above that tweet count for a day can be about 500 million. These tweets may contain lots of information regarding stock market prices.

This thesis makes use of twitter for obtaining public views or as a news source because:

* Buying a news API will undoubtedly cost a lot of money. At the same time, twitter provides it for free.
* Now, most of the people make use of social media for obtaining news (Anon., 2016).
* Using Twitter, one can get most up to date news.
* All major news company uses twitter account for giving up to date information.
* In the 21st century, social media has become the best pathway for obtaining live news.

The information will be fed to different machine learning algorithm along with an ensemble algorithm, and a comparative study of the general machine learning algorithm and an ensemble-based algorithm will be performed. The idea here is to boost the overall accuracy of the system using a hybrid algorithm.

* **Ensemble-based algorithm** (Lutins, 2017)**:**

Ensemble machine learning algorithms are actually hybrid machine learning algorithms that combine multiple machine learning algorithm in some pattern to overcome their individual frailty and boost the overall accuracy. Types of Ensemble algorithm are:

1. **Bagging** (Dey, 2019)**:**

Bagging consists of multiple weak learning algorithms that perform learning independently and parallelly. These are homogeneous weak learners. The output of all the learners is combined in the end to form a final outcome. Some kind of deterministic averaging process is followed for obtaining this final output.

1. **Boosting** (RAY, 2015)**:**

Boosting also consist of homogeneous weak learners. Here learning does not take place parallelly but in sequential order. Each model output is forwarded as input for the next model in the iteration.

1. **Stacking** (Smolyakov, 2017)**:**

Learners in stacking are heterogeneous and weak in nature. Learning in stacking happens parallelly. All weak learners perform learning in a parallel manner, and then they're out is combined and given as input to a meta classifier also known as the final learner—the output obtained by this meta classifier id the final output.

## 1.3 Problem Statement:

Investing money in share market is more of gambling nowadays. It is challenging to predict the share prices of a company. People have gone bankrupt to losses in share market. Researchers are continuously researching to analyze some pattern in share prices, but it is challenging to find anyway. Some have also used social media websites to gain up to date information about share market through public views. But the accuracy obtained was not very significant.

**1**.4 Goals:

Sentimental analysis of tweeter will bring us excellent knowledge about the stock of the company and there up and down during a phase. Using the sentimental analysis, we can predict the stock price of the stock. People view to determine the company reputation, which affects the company, which leads to the stock market and change the price of the stock of a particular company. So, one of the biggest reasons for choosing this topic is the same.

**Research Question:**

* This thesis has the aim of conducting a comparative study of machine learning algorithms to determine which algorithm can perform better with twitter data sentiment analysis.
* The thesis will make use of Neural network-based machine learning algorithm, in order to see if neural networks can produce better accuracy when compared to general machine learning algorithms.

## 1.5 Structure of the document:

* **Literature Review:**

This section will discuss the research paper studied up until now by me to come up with this project. It will discuss the past methodologies used by different researchers to perform sentiment analysis on big data. An analysis will be done on the advantages and disadvantages of those methodologies. At last, the section will consist of the entire literature review conclusion.

* **Proposed Methodology:**

This section of the proposal will discuss the algorithms that will be used for the development of this project. A step by step implementation of the algorithms will be discussed. At last, the section will also consist of the system architecture of the project.

* **Implementation:**

This section will discuss different libraries that will be used for the implementation of this project. It will consist of a flow diagram that will explain the full implementation of the project in detail.

* **Proposed Evaluation:**

This section will consist of the evaluation measures using which I am going to evaluate the results obtained by the implementation of the project. Each evaluation measure will be discussed in detail.

* **Conclusion:**

This section will conclude the entire proposal. A concluding statement will be given here regarding the implementation of the project.

* **Reference:**

This section will consist of the list of papers, articles, etc. used by me for coming up with this project.

# . Background

## 2.1 Introduction:

This section will discuss the literature studied by me for the development of this thesis. A discussion will be done on different methods used by various researchers for their research. We will discuss the pros and cons of other studies, and at the end literature review, the discussion will be done.

## 2.2 Literature Review:

* **Machine Learning** (Anon., 2020)**:**

The learning of these models by algorithms are basically of two types:

1. Supervised Learning
2. Un-Supervised Learning

This thesis makes use of supervised classification-based machine learning algorithms for predicting future stock values.

* **Supervised Learning** (Edwards, 2018)**:**

In Supervised Learning, the algorithms build a mathematically based model based on the input dataset given to it. Here algorithm has both input and output inside the dataset. A form of learning that contains both input and output data as the dataset is termed as Supervised Learning. Here, for example, if the algorithm has to figure out if the email is spam or not. These 0 and 1 based Boolean problem comes under supervised learning.

In a broad sense, the algorithms that come under Supervised Learning are Classification algorithms and regression.

If the output for some dataset is definite than classification algorithms tends to perform well as compared to other machine learning algorithm. For example, we are detecting whether an email is spam or not.

* **Regression** algorithms are used when the output for actual data is limited or restricted within a range but in a continuous manner. For example, the continuous value of change in temperature.
* **Ensemble algorithms** (Lutins, 2017)**:**

Several models are used in combination in order to improve the final accuracy in case of an ensemble-based approach. Hence, an ensemble algorithm is found to produce better results as compared to standard algorithms.

With the aim of improving the overall prediction accuracy and decrease the variance and bias, a model is built with a combination of various machine learning algorithms. The concept is known as the ensemble method.

Ensemble method can be categorized into two main divisions:

1. Sequential based ensemble models
2. Parallel based ensemble models

Sequential based ensemble models are models where learning takes place sequentially. Parallel based ensemble method consists of models where learning takes place parallelly and independently.

## 2.3 Twitter Overview:

**(Doshi , et al., 2017)** Twitter is a social media app, also known as a micro-blogging application whose popularity has grown far across the world. Twitter keeps people socially connect to the world. With the help of twitter, people can view live updates about the tweets whom people follow and can also write their own tweets. Twitter contains up to date information, be it a job vacancy or an earthquake in a particular place. Stored tweets on twitter can be considered as an unstructured form of data that can be used for obtaining information regarding a matter.

Twitter is, of course, one of the most popular social media applications on which the altmetrics are founded. According to **(PARKER, 2020)** the most respected financial journals across the world have a twitter feed. There is various feeds provider online that can help investors to make the proper decision regarding the stock market.

For example, CNBC provides real-time business information regarding on worldwide basis. This coverage is for about 15 hrs and is focused towards all the individual investors.

@Benzinga proves real-time financial news and in-depth analysis and breakdown of live share market news.

(**Mankar, et al., 2018)** proposed a system to make a stock market prediction based on machine learning algorithms. A plethora of real-world problems is being solved using data mining and machine learning approach (Mankar, et al., 2018)**.** The techniques described were found to be significantly effective, which resulted in yielding max results with very little investment and also was proven to be time-efficient. Stock markets are the most lucrative option for people to add some extra money to their annual income.NLTK library was used for removing stop words and other pre-processing tasks for building a machine learning-based model.

After the pre-processing of the gathered data from twitter, the machine learning algorithms are applied to the classifiers. Naïve Bayes and Support Vector Machine are the two algorithms used for extracting features from a list of elements present in the dataset. Feature matrix was created using unigram tweets. The newly generated dataset with a new set of features is subjected for training with classifiers like SVM and Naïve Bayes algorithm. The sentiment for each tweet was extracted into three sections as positive, negative and neutral. On the other side, they are Taking the historical stock market data every day, calculating the market up and down and as a dataset label. Historical data for the stock market, Python was used as a yahoo-finance library. The results obtained stated that SVM proved to be much more efficient with dealing with the stock market.

* **Drawbacks:**(Mankar, et al., 2018) The accuracy obtained was not very good. The system lacked in giving high performance. Inclusion of multiple moods while building model with NLP lead to a decrease in overall accuracy. The dataset used was very small.

(**Pagolu, et al., 2016)** Developed a system to predict stock market movements. They conducted their research based on (Fama, n.d.). According to (Pagolu, et al., 2016)the prices involved with stocks in the stock market fluctuates a lot. Studies conducted by researchers have debunked the myth of predicting stock market with the historical help data. They found that financial market prices largely depend upon news, social media and current event. During pre-processing, the missing value was filled with using the approximation function. Methods like tokenization, Stop words removal were performed. The pre-processed data set is trained using Random Forest classifier. The results obtained showed that a strong correlation exists between the rise and fall of the stock prices of any company to the public opinion and emotions about the company expressed in tweets.

(**Mittal & Goel, 2012)** Developed a system to predict stock market movement using machine learning algorithms. Twitter was used as a dataset, and DJAI value was used to predict the stock market movement. EMH (Efficient Market Hypothesis) indicates that stock prices are primarily driven by information spread on social media. This paper tests the hypothesis that emotion and mood of the people can have a direct effect on stock market prices. The DJIA value dataset is first processed. The tweets obtained are then forwarded to algorithms related to sentiment analysis. The results obtained by the functioning are divided into four classes. The tweets obtained, along with processed DJIA value, are fed to the machine learning model. The accuracy received was about 75%.

(**Vu, et al., 2016**) performed an investigation on CNN and RNN performance for relation classification. According to the results obtained by their research, the performance of CNN was observed to be better than RNN. They both resulted in entirely complementary information. The results showed that the RNN provides more of a weighted combination of words present in the sentence while CNN is based on the function of extracting n-grams. (Yih & Richardson, 2016) and (Adel & Schutze, 2017) also supported CNN over other NLP based algorithms like GRU or LSTM for classification of huge sentences.

In 2017 (**Yin, et al., 2017)** conducted research to perform a comparative study between CNN and RNN algorithms. Seven different types of tests were shown.

Sentiment Classification (SentiC): The test was performed with SST dataset **(Socher, et al., 2013**). The dataset was related to movie sentiment analysis based on the negative and positive reviews given. After splitting machine learning algorithms were applied, and accuracy was obtained. The labelled phrases were treated as independent training instances. Accuracy was obtained for both CNN and RNN.

Relation Classification (RC): **(Hendricks et al., 2009)** data set was used for performing relation classification. F1-score was obtained for the algorithms. The dataset consisted of sentences which were manually labelled. The training dataset consisted of 8000 sentences, while the testing dataset consisted of 2717 sentences.

Textual Entailment (TE): The dataset was obtained from (Bowman, et al., 2015). The sentences involved in the dataset were premise-hypothesis pairs, labelled with a relation (entailment, contradiction, neutral). The dataset was split into train and test after removing unlabelled pair. The accuracy was measured for both CNN and KNN.

Answer Selection (AS): The dataset used consisted of an of open domain question answer. Two more kinds of the dataset were used for obtaining an accuracy comparison of CNN and RNN, which were Question Relation Match (QRM), Path Query Answering (PQA) and Part-of-Speech Tagging. By performing these tests **(Yin, et al., 2017)** were able to deeply understand the difference in the working of RNN and CNN with respect to NLP.

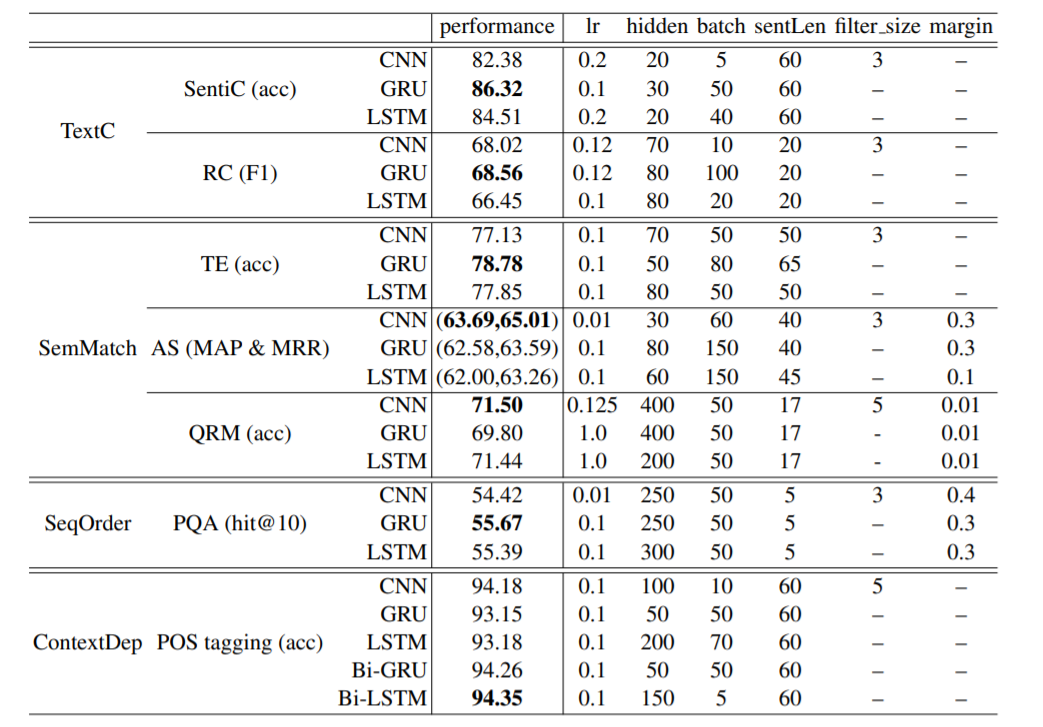


Table . (**Yin, et al., 2017)** Performance report of CNN and other KNN algorithms

Table 2.1 displays the results obtained by the experiments conducted with CNN and KNN algorithms with respect to all the hyperparameters. In terms of SentiC GRU, a KNN based algorithm performs best. CNN gives the best performance when it comes to AS and QRM. GRU serves better than CNN and LSTM when it comes to TE. In the case of SeqOrder and ContextDep, CNN has the worst performance. The results for ContextDep and SeqOrder are as expected: RNNs are very much suited for encoding order information (for PQA) and long-range context dependency (for POS tagging).The overall conclusion made was that the RNN outperforms CNN in a wide range of task except when it comes to key phrase detection and question-answer based sentences.

(**Lei & Wang, 2015)** developed a system to predict the price trend in the stock market using data mining algorithms. They first made use of anomalies to predict the upward trend in the stock market on the distribution of trading prices volume. Tick by tick data was used on stock prices instead of the time series. The experiment was conducted by selecting 200 stocks randomly. According to the results obtained, it was clear the upward trend in the stock prices can be predicted using anomalies. The dataset in this thesis consisted of tick-by-tick format. This data records each trade for every stock in the market, if there are 1000 trades for a specific stock then there will be 1000 records for that stock on that day, so for a relatively long period, the data size can be massive. One record of the tick-by-tick data is defined as:

R = {p, t, v, c, a, b}

Equation .

An outlier mining algorithm was proposed for observing the trend in the stock market.

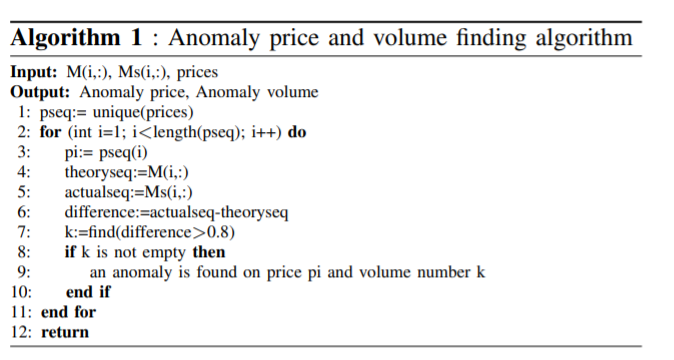


Figure . **(Lei & Wang, 2015)**

A = {s, p, v} describes an anomaly record where stock index = s, anomaly price = p, anomaly volume = v. Another algorithm is used after all the anomalies get defined in order to locate those anomalies on time during the trade.

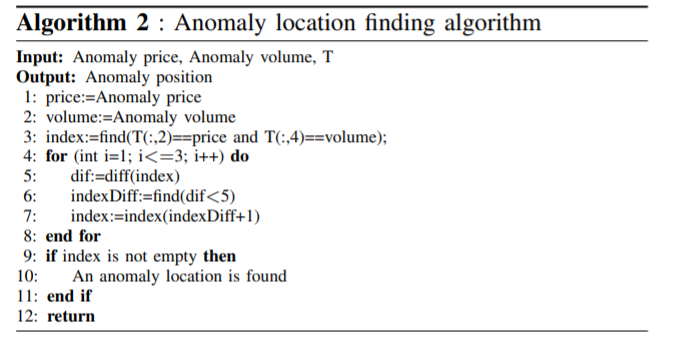
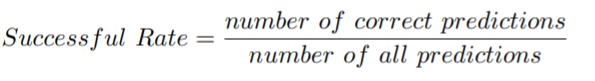


Figure .(**Lei & Wang, 2015)**

From Chinese Shenzhen stock market, 200 stocks were selected randomly for the research. It was found that 111 displayed the behaviour of a cluster of anomalies. A graph was plotted based on the average return of stock after anomaly for 100 days. Same stocks were also measure using the SVM algorithm for comparison purpose.



Equation .

The results of the experiment conducted showed that the algorithm proposed in this thesis outperforms the SVM results. The proposed algorithm outperforms traditional data mining algorithms. The experiment data are from Chinese stock exchange with the time range 03-31-2014 to 04-30-2015, which include 272 trading days. The size of the data set is 7.1 GB.

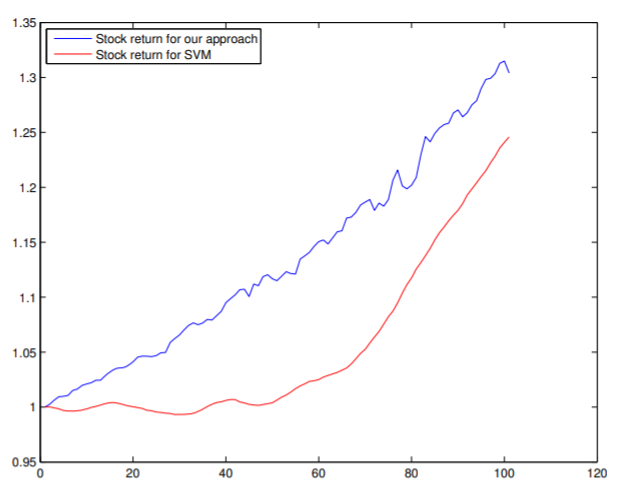


Figure .(**Lei & Wang, 2015)**

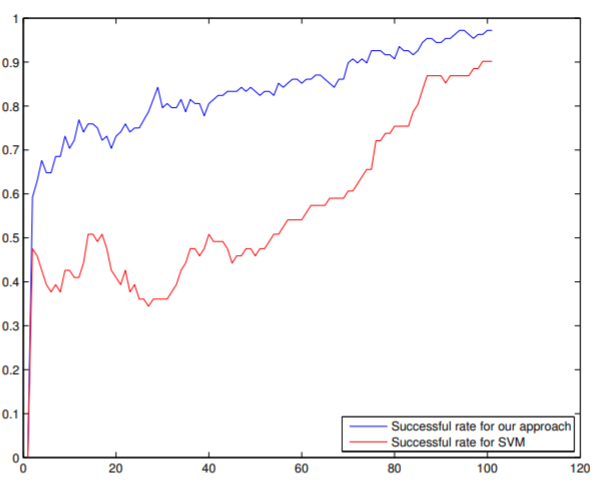


Figure .(**Lei & Wang, 2015)**

* **CNN (Convolution Neural Network):**

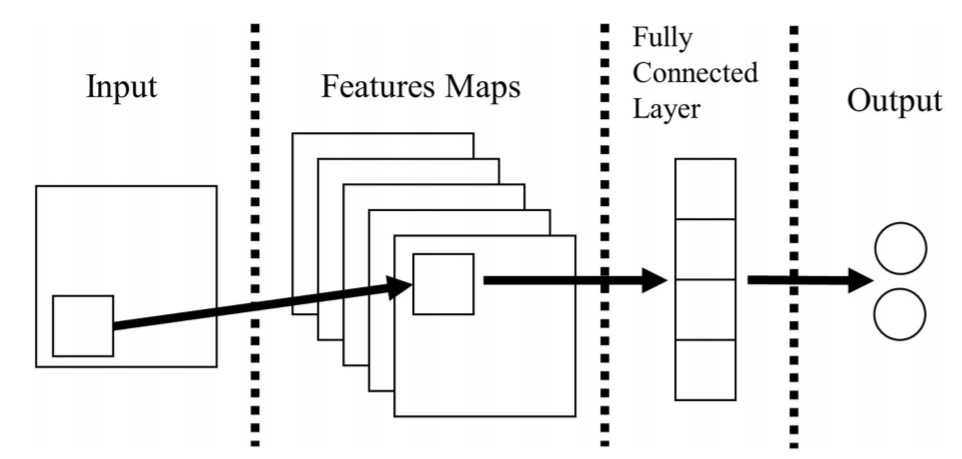


Figure . (Jang, et al., 2019) general CNN model

Fig represents a model of CNN algorithm in very brief terms. CNN algorithm is for classification of the features and extraction. No information losses occur while passing of values from one layer to another. The semantic similarity between words can be gained using spatial information with the help of CNN algorithm. There are multiple hidden and output layers present in CNN method. Various convolution and pooling layers are present in hidden layers. A feature map is also present for each layer. The work of the convolution and pooling layer is to get the input value characteristics and map them on to the feature maps. With these characteristics, the semantic between the words of a sentence can be extracted. Steps involved in CNN are:

* The parsed data from the input layer is passed on to the feature maps.
* The convolution operations are performed, and the data is mapped on to the pooling layer.
* A max-pooling operation is performed on the data before it gets mapped.
* The most massive value is extracted from the previous results during max pooling operation.
* A fully connected layer gets created by the CNN subsequently. All the convolution and pooling layer gets combined over here.
* The results finally get outputted to the final layer, which is the output layer.

**Word2vec:**

The CNN can be used with oneHotEncoder or with word2vec based vectors. This thesis will make use of Word2vec for generating vectors. Word2vec makes use of the hypothesis that words with the same meaning tend to have close distance in the document. Using the hypothesis, Word2vec generates vectorises the words meaning present in the sentence.

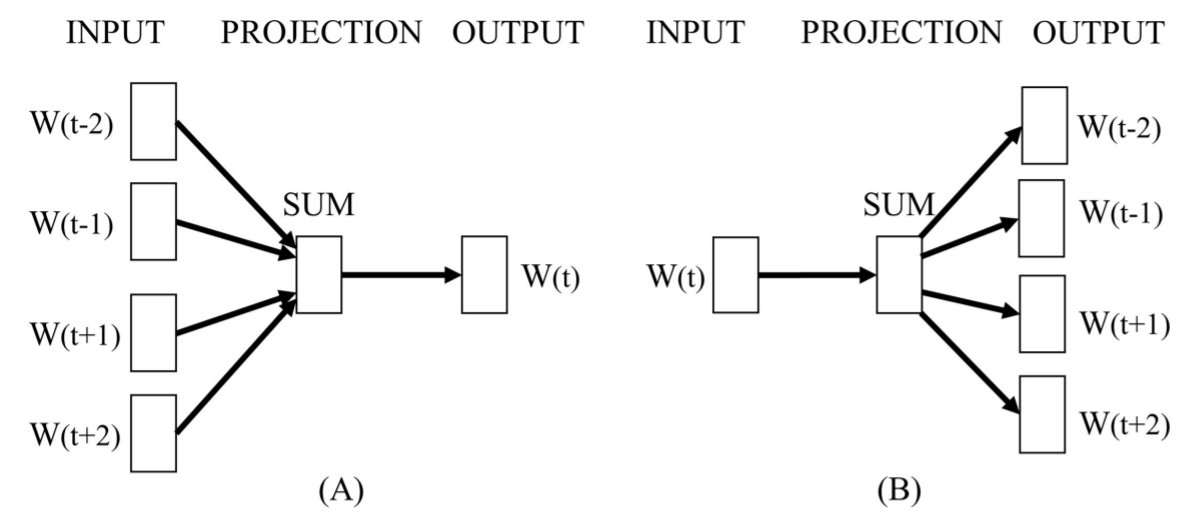


Figure . (Jang, et al., 2019)

(Jang, et al., 2019) developed a system for classification of news articles based on twitter sentiment analysis using Convolution Neural Network algorithm. Word2vec was used for classification model generation. The research was also focused on analyzing the effects of Word2vec on the results. The results showed that the accuracy of a classification model is boosted significantly when word2vec is used for learning the semantic relation of words. They concluded that CNN method could perform faster and more accurate classification when the data provided to it is in a formatted manner. The results were measured in terms of accuracy, recall and f1-score.

## 2.4 Related Work:

As discussed by (Jang, et al., 2019) CNN can use both OnehotEncode and Wordtovec methods for generation of vectors. According to the results obtained by (Jang, et al., 2019) wordtovec boosts the overall accuracy of the system as compared to the onehotencoder. This thesis will make use of wordtovec for generating vectors.

(Yin, et al., 2017) conducted a comparative study of both KNN and CNN algorithms. The results showed that the RNN outperforms CNN in a wide range of task except when it comes to key-phrase detection and question-answer based sentences which is precisely the kind of work that will get performed in this thesis. Hence, the thesis will use CNN algorithm for performing sentiment analysis.

# . Methodology

## 3.1 Overview:

* **CNN algorithm** (Tran, 2020)**:**

A neural network is a machine learning approach which functions by finding out any existing patterns from the given input vectors or numbers. The vectors or numbers are created by transforming the real-world data by encoding them in number format into the vectors. CNN can be described as a neural network-based algorithm that **applies convolutional layers to local features.**

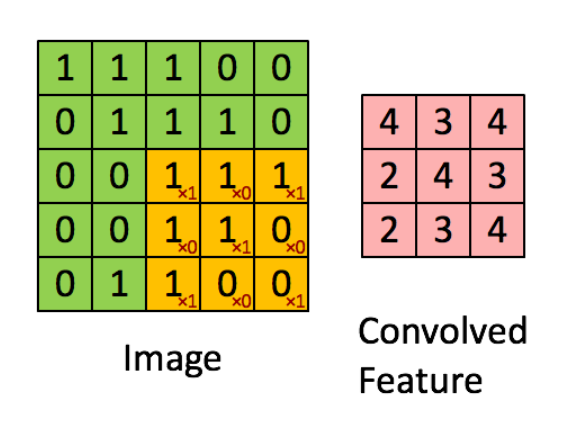


Figure . (BRITZ, 2015) matrix and filter

Let us consider the left matrix as an image with only black and white color. Each instance in the matrix, i.e. 1 and 0 represents one pixel of the picture. 0 represents a black pixel, while 1 represents a white pixel. Consider the right-side matrix as the sliding window it is also termed as kthe ernel, feature detector or filter. The filter detector is a 3\*3 matrix. Each value of this kernel is multiplied with the values of the original matrix. The final answer to this is the sum of all these multiplied values. To obtain value for all index of matrix or to get a complete convolution, the matrix is moved across the entire image matrix in a sliding window manner. These filters can also be used with an image having more color than black and white.

* **NLP (Natural Language Processing) with CNN (Convolution Neural Network)** (BRITZ, 2015)**:**

In the case of images, pixels are given as input which is then transformed into vectors. But for NLP sentences and documents are converted into a matrix. Each row in the matrix represents a token value. Which means each vector is used to describe a word or a character. This thesis makes use of word2vec function for encoding a given sentence into vectors. If a sentence consists of 10 words, then the matrix formed will be of size 10\*100. The matrix formed will be similar to a matrix that is created from an image. In case of an image, the filter slides over the image, while in case of NLP the filter slides of the sentence.

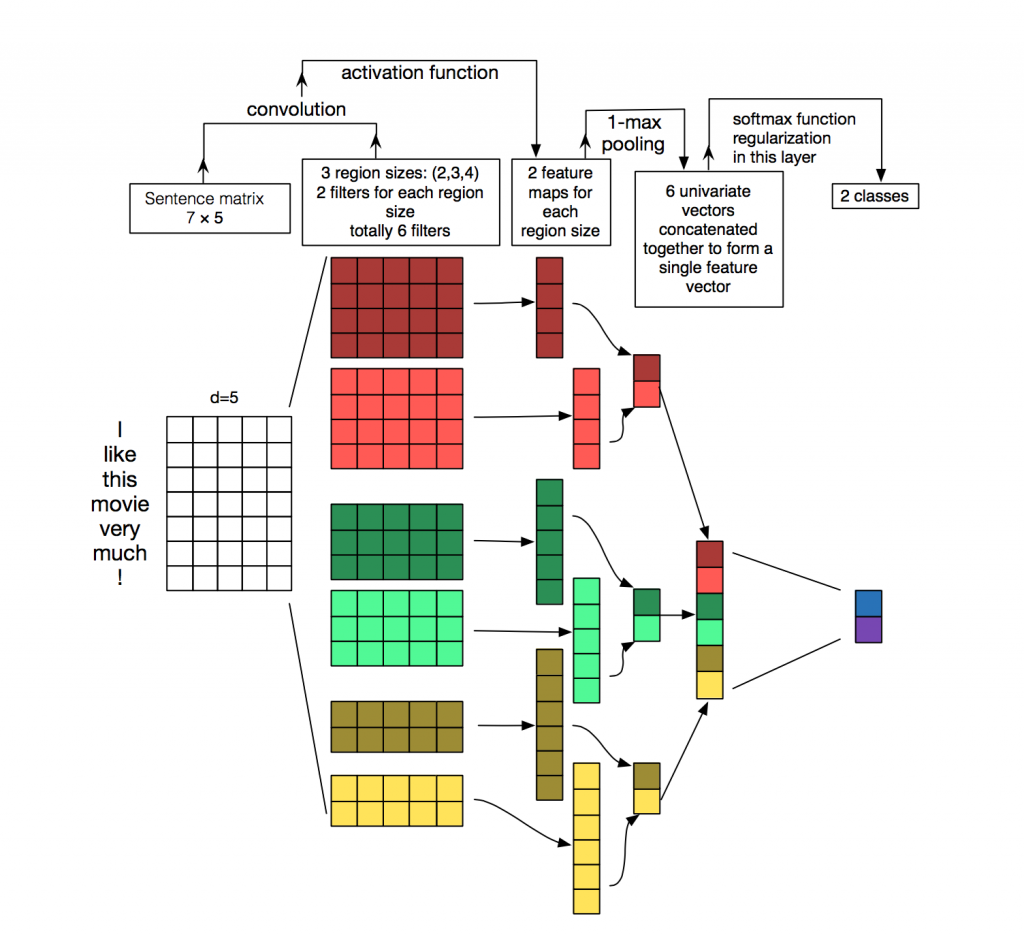


Figure . (BRITZ, 2015) CNN architecture for NLP

Fig 3.2 displays the architecture of CNN when used for Natural Language Processing. A number of filter regions are three of sizes 2, 3 and 4. Each is having a set of filter. As soon as convolution is performed on the sentence, matrix feature maps are generated. Each feature map has the largest number obtained by performing a poll for each feature map. These maximum numbers obtained together creates a univariate vector based on features. The six maximum feature value for the penultimate layer will be combined to produce a feature vector. The final layer of softmax receives this vector input and uses it for the classification of The sentence; here we are assuming the binary classification and thus represent two possible output states.

* **CNN vs KNN (BRITZ, 2015):**

One of the biggest advantages of CNN over KNN is that CNN is fast. Very fast. The central part of any computer graphic is convolution. CNN is implemented over hardware level on GPUs. When it comes to the representation part, CNN is far more efficient than n-grams. Computing above 3-grams can be quite an overhead for the system with the availability of large vocabulary. Convolutional Filters has the ability to learn good representations automatically,without facing any need to represent the entire vocabulary.Having filters of size greater than 5 is a completely reasonable option. I like to think that there are many of the filters learned in the first layer are quite similar (but not limited) to n-grams and are compact.

* **Hyper Parameter available in CNN:**

The application of a 3\*3 filter at the center of the matrix is simple. But in order to apply a filter to each and every element available in the matrix, the concept of zero paddings is used. Elements outside the matrix are taken as zero. This helps in applying a filter to each element available in the matrix.

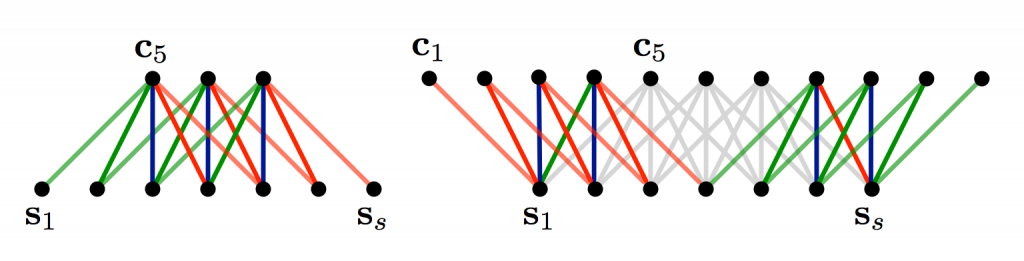


Figure .(BRITZ, 2015) Narrow and wide convolution

From fig 3.3, it can be seen that how useful can wide convolution can be, or even necessary when we have a filter relative to the large input size.In the above, the narrow convolution yields an output of size

**(7-5) + 1=3**

Equation .

**,** and a wide convolution output of size

**(7+2*4 - 5) + 1 =11**

Equation .

**.** More generally, the formula for the output size is

**n_{out}=(n_{in} + 2*n_{padding} - n_{filter}) + 1 **

Equation .

* **Stride Size:**

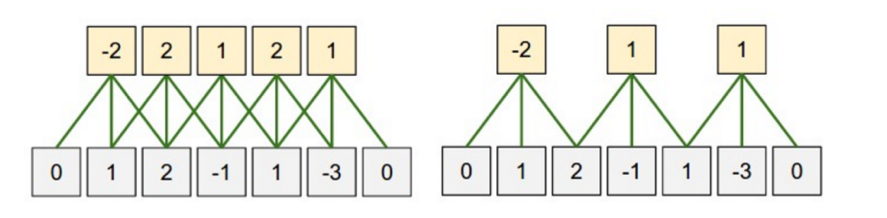


Figure . (BRITZ, 2015)

Stride size is another hyper-parameter used in CNN. It defines the distance with which the filter must be shifted during each step. The standard Stride size used is 1. A filter with larger size leads to smaller size outputs due to very few application of the filter.

#### **Pooling Layers**

(BRITZ, 2015) Pooling layers are one of the key aspects for CNN algorithm. Its application happens just after the convolution. The main objective of pooling is to subsample the input provided. A maa operation is done on the result of each filter with the help of pooling. Pooling over the entire matrix is not required. For example, the following shows max pooling for a 2×2 window (in NLP we typically are apply pooling over the complete output, yielding just a single number for each filter):

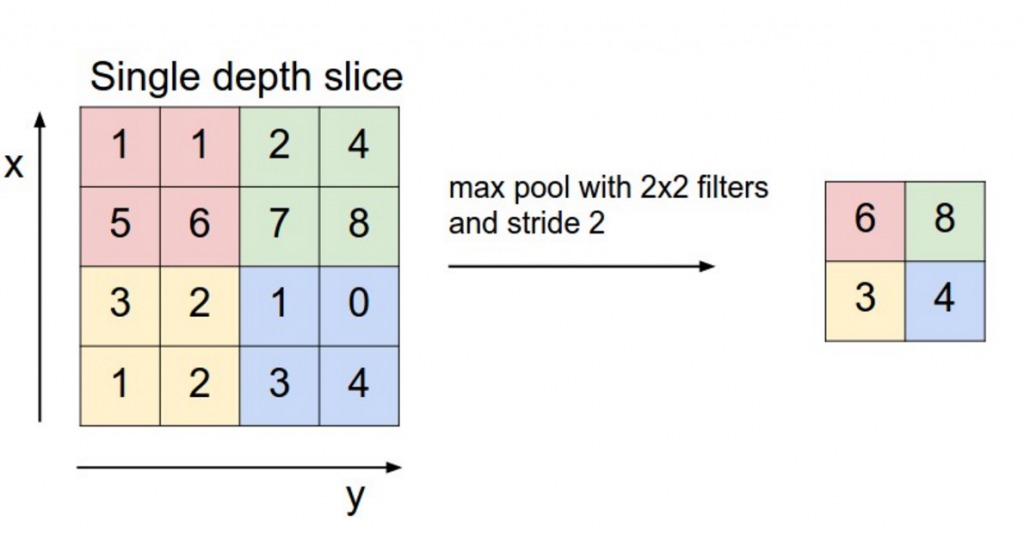


Figure . (BRITZ, 2015)

Fig 3.5 displays maximum pooling using a 2\*2 filter with stride 2. The text is in the form of a 4\*4 matrix. A 2\*2 filter is applied.

**Why Pooling is used** (BRITZ, 2015)**:**

There are basically two properties of pooling: It has the ability to generate a fixed size matrix as output in case of classification.

Example: consider a scenario of having one thousand filters. The application of pooling to each of these filters will result in an output of 1000-d irrespective of the filter size or input size. This gives the flexibility of using sentences and filters with variable size and get similar output-D as an input to the classifier.

One of the advantages of pooling is that it removes the overall output dimensionality without removing the most salient features. A specific feature is detected by each filter. For example, if there exists a negation in the sentence like “not amazing”. If such kind of negation is found at some other place in the sentence, then the application of the filter in the area will result in a large value but in other regions a small value. The application of the max operation helps to keep track of features appear in the sentence while in-turn losing the ability to track the emergence of feature location in the sentence. The question of whether the locality of feature emergence information helpful or not. It can be said that it is helpful, but it is quite similar to what the bag of n-grams model is doing. The outcome is that we are keeping track of local-level information using filters while losing the ability to track global level information regarding locality.

# . System Design and Specifications

**System Specification:**

This section will describe all software and hardware level tools used for developing this thesis.

## 4.1 Software Specification:

|  |  |
| --- | --- |
| Operating System | Windows |
| Language | Python |
| Code Editor | Visual Studio Code |

Table .Software specification

Table 2 denotes the software level tools used for developing the thesis. As mentioned in the table, windows are used as the operating system. Python is used as a base language for writing all machine learning code. VSCode is used as a base code editor for writing python code.

## 4.2 Hardware Configuration:

|  |  |
| --- | --- |
| RAM | 4GB |
| HDD | 10GB |
| Processor | Intel 1.66 GHz |

Table . Hardware Configuration

Table 3 displays the hardware level tools used for developing the system.

## 4.3 Libraries used were:

Below mentioned is the list of libraries used for writing machine learning code.

* **NumPy** (Anon., 2020)**:**

Arrays are implemented in python with the help of NumPy. Various operations like linear algebra, matrices and Fourier transform are performed with the help of functions existing in NumPy. Usually, developers make use of list wherever they can instead of going with NumPy based array. But the list cannot be used in place of arrays every time.

**Advantages of NumPy over lists** (Anon., 2020)**:**

The subsequent memory locations are occupied by elements in the NumPy array. This subsequent placement of elements of an array in NumPy makes accessing the element a lot faster than a list. This concept is also well known as the locality of reference. NumPy is also optimized to work with the latest CPU architecture. It is also an open-source library with python-support (Anon., 2020).

* **Pandas:**

Pandas, in general, is an open-source library or tool available in python providing high-performance tool. It is easy-to-use and issued for dataset manipulation and data analysis. A variety of industries worldwide make use of pandas which include economics, Statistics, analytics, etc.

* **Sklearn** (JAIN, 2015)**:**

It is a machine learning library an open-source package available in python. It consists of various supervised and un-supervised form of machine learning algorithms bundled together. Different types of machine learning algorithms like classification, regression, dimensionality reduction and clustering are present in this package.

Components of Sklearn are (JAIN, 2015):

Supervised and unsupervised learning algorithms, Cross-validation, Feature extraction, various toy datasets.

* **Keras** (Heller, 2019):

Keras is a super friendly deep learning open source library. Apart from being user friendly and ease of learning Keras has other advantages like broad adoption and have support from a variety of deployment options with a minimum of 5 backend engines, which include TensorFlow, CNTK, Theano, PlaidML and MXNet. It has strong support for multiple GPUs. Keras is being backed by a lot of tech giants like Amazon, Uber, Apple, Google and Microsoft.

* **Tkinter** (Anon., n.d.)**:**

Tkinter is an open-source GUI library used in python. Using Tkinter in python fast and decent GUIs can be developed. Tkinter is based on object-oriented programming approach.

Steps to create a GUI with Tkinter:

* Importing the module
* Design a main window GUI app.
* Insert the necessary widget to the GUI app.
* Write the necessary actions to perform while a particular event is triggered.

Tkinter contains various widgets like Button, Check Button, Canvas, Frame, Label, Menu Button, Menu message, Scrollbar, Text, tkMessagebox, PanedWindow etc.

* **Design:**

The design section will discuss the flow of information for both the training and testing phase using flow diagrams.

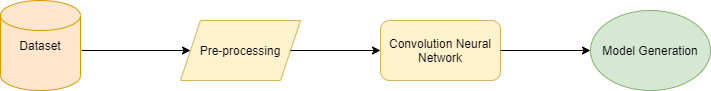


Figure .Training Phase flow diagram

Fig represents a flow diagram for the training phase. The dataset consists of tweets from twitter. Pre-processing is done on the available input dataset. Various operations are performed during the pre-processing stage.

For e.g.:

**Normalizing of words:** Converting all the available words into lower case. Keeping a common format for all the available inputs.

**Cleaning of data:** It involves removing all alpha-numeric values which have no contribution towards the output. Trimming of words, removing stop words, removing null values, removing garbage values etc.

After the data is being cleaned and pre-processed. The adequate machine learning algorithm is applied to it, in order to build the model. For this thesis, CNN algorithm is used for the development of the model.

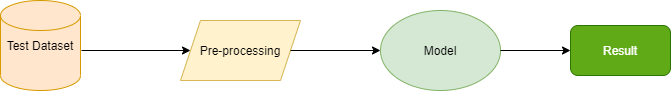


Figure .Testing flow diagram

The testing dataset is first subjected to pre-processing. Pre-processing removes all unwanted features and data which does not have any impact on the final prediction. After the test data is pre-processed, it is inputted to the model developed during the training phase. Based on the training of the model and the test data inputted, the model makes a prediction.

# . Implementation

## 5.1 Architecture:

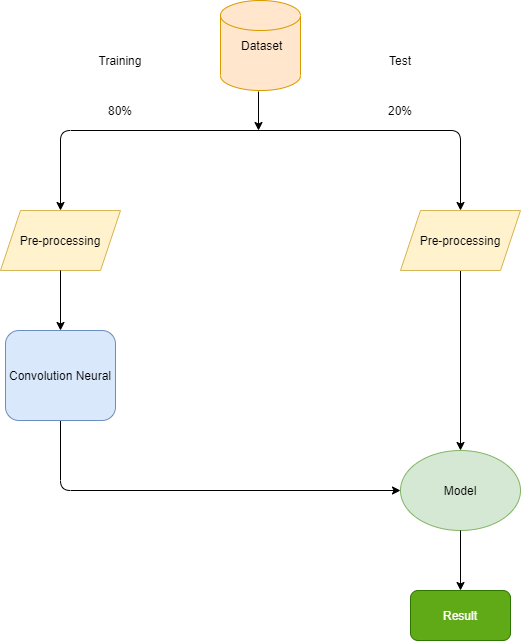


Figure . Architecture

The dataset is split into two parts. 80% of the dataset is used for training the model, while 20% is used for testing purpose. As discussed in the design section, the training data is first pre-processed. The pre-processed data is then subjected to the CNN algorithm, which in turn leads to the formation of a machine learning model.

The testing dataset is fed to this model for getting the results.

## 5.2 Data pre-processing:

Data pre-processing is the first step after the loading of the dataset. Data pre-processing involves removing all the garbage values and null values from the dataset. The words from the sentence get tokenized for sentiment analysis. Removal of stop words is done. Stop words are general grammatical words which do not carry any meaning with them. Normalization, tokenization and cleaning of words are done.

## 5.3 Data Visualization:

Seaborn, matplotlib and pandas libraries are used for performing data visualization. Representing information on machine learning model in the form of graphs is the work of data visualization.

## 5.4 CNN based Model:

Finally, after all, pre-processing steps on the dataset. The modified dataset is then fed to the CNN algorithm for the development of semantic analysis model.

## 5.5 User Interface:

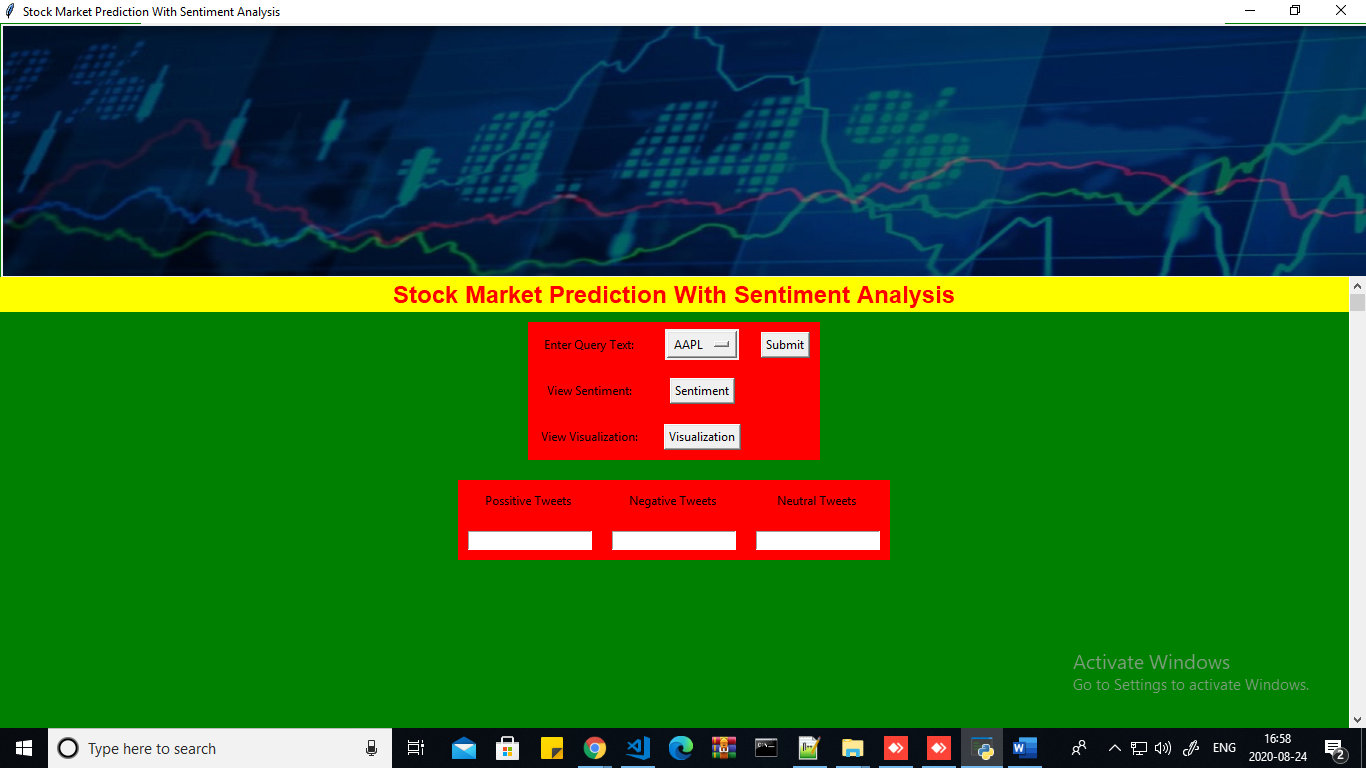


Figure . UI design

Fig 5.2 displays the GUI of the software developed by us. The first option indicates the selection of the company for which we want to know the sentiments. The checkboxes below display the sentiments of the company. The checkbox with Positive label tweets will display the percentage of the positive tweet about the company on twitter. The checkbox with a negative tweet will display the number of the negative tweet about the company. At the same time, the checkbox with neutral tweet will display the number of the neutral tweet about the company.

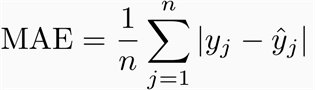
# . Results

## 6.1 Introduction:

The result section will discuss about the results obtained after the implementation of Convolution Neural Network Algorithm. It will also compare the results with previously obtained results in the literature review, and finally, a conclusion will be made:

1. (MAE) Mean Absolute Error
2. (MSE) Mean Squared Error
3. R2 score

* **Mean Absolute Error (MAE):** MAE is the measure of the error in terms of magnitude for a set of predictions. The results obtained has no data regarding the error direction it only contains magnitude. In simple terms, it is the average error value obtained between the prediction made and the actual observations. Predictions are considered as equal weights in terms of MAE.



Equation .

* **Mean Squared Error (MSE):**

MSE or Mean Squared Error is the average measure of the square of the error rate produced due to the difference in the predicted results and the actual results. It is also known as a risk function as it directly corresponds to the expected value of error loss squared. It incorporates the variance along with estimators as its bias.

* **Steps to find the MSE:**

1. Obtain an eq for the regression algorithm line.



Equation .

1. Substitute the value of “x” obtained by the implementation of the first step to get the value of Y.



Equation .

1. The new Y value obtained is the predicted value. So this Y value has to be subtracted from the actual value to obtain the error rate. This will generate a value which is the actual error term. Also denotes the verticle distancing from the line of regression.



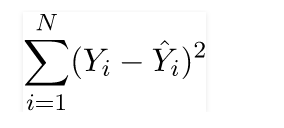
Equation .

1. The error value found from step 3 is squared now.



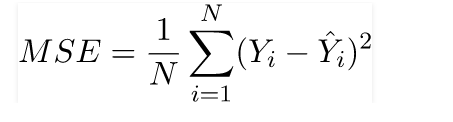
Equation .

1. A summation is obtained for all the error values.



Equation .

1. The final value obtained in step 5 is divided by the total number of observations.



Equation .

* **R2-Score:**

[R-squared](https://statisticsbyjim.com/glossary/r-squared/) actually denotes how well a model is fitted to the given data, also known as the goodness of fit. It also gives the measure of the percentage of variance with the dependent and independent variables. R-square in overall terms is used to measure the strength of the relationship between the developed model and the dependent variables. The measure ranges on a scale from 0-100%.

The data points scattered around the regression line is evaluated by R-square, also known as the determination coefficient for regression. The higher the value of the R-square measure, the smaller is the error value or the difference between the actual model and the predicted one.

{\displaystyle R^2 = \frac {\text{Variance explained by the model}}{\text{Total variance}}}

Equation .

The value of R-square always lies between 0 and 100%.

## 6.2 Results:

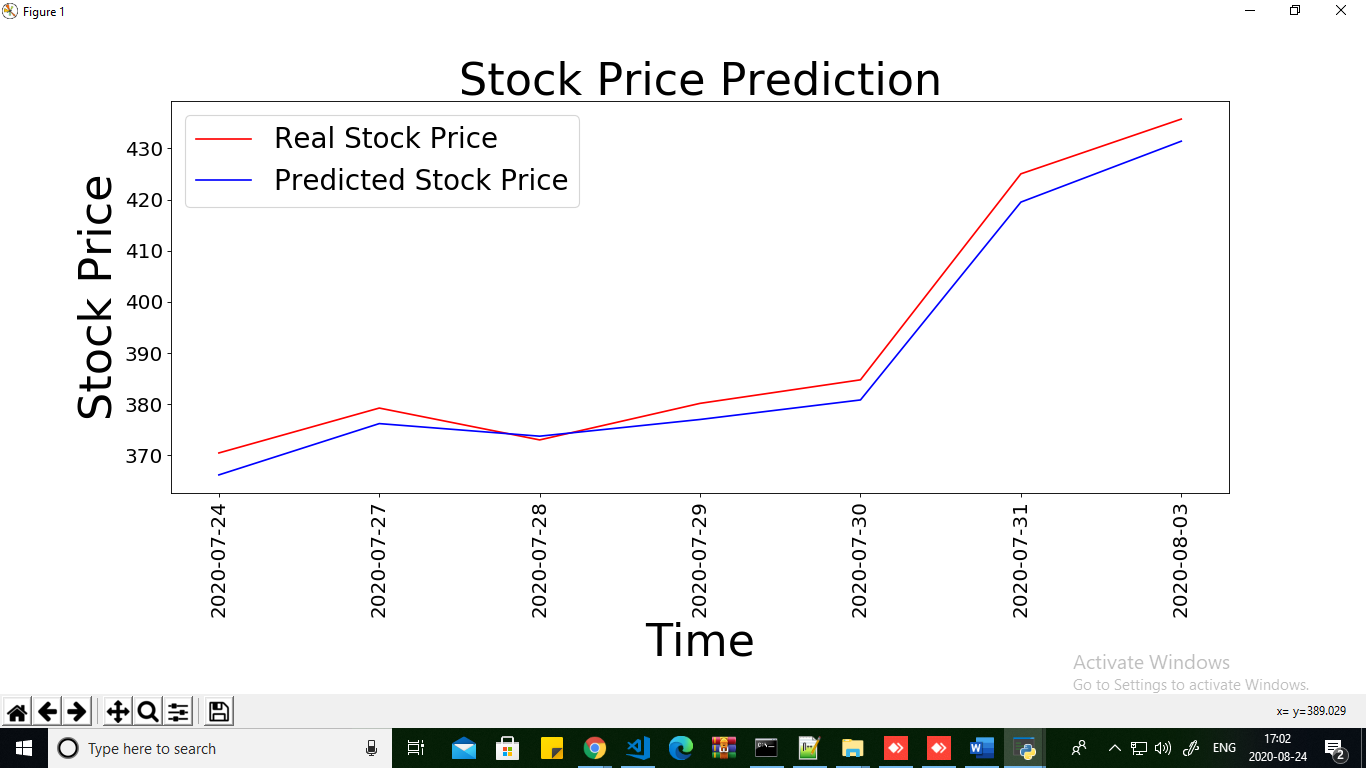


Figure . Prediction result

Fig 6.1 displays a graph for stock price prediction. The graph shows the relation between stock price and Time. This graph will be obtained with the help of the visualization button on the GUI.

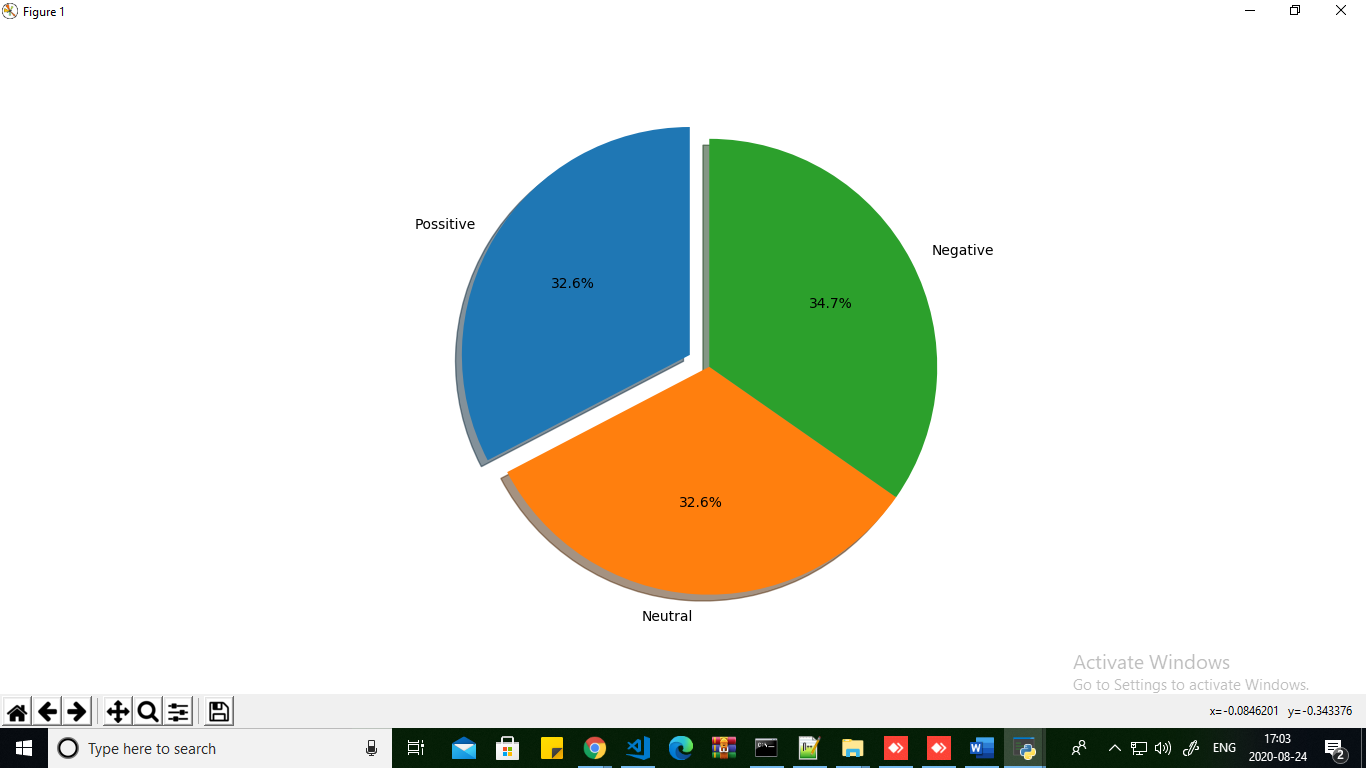


Figure . Sentiment result

Fig 6.2 represents the sentiment of tweets obtained regarding the company we selected in the sections of positive, negative and neutral.

* **R2-Score:**

[R-squared](https://statisticsbyjim.com/glossary/r-squared/) actually denotes how well a model is fitted to the given data, also known as the goodness of fit. It also gives the measure of the percentage of variance with the dependent and independent variables. R-square in overall terms is used to measure the strength of the relationship between the developed model and the dependent variables. The measure ranges on a scale from 0-100%.

The data points scattered around the regression line is evaluated by R-square, also known as the determination coefficient for regression. The higher the value of the R-square measure, the smaller is the error value or the difference between the actual model and the predicted one.

{\displaystyle R^2 = \frac {\text{Variance explained by the model}}{\text{Total variance}}}

Equation .

The value of R-square always lies between 0 and 100%.

The results achieved after the implementation of the model is:

**The r2 score obtained is: 0.9612**

# . Conclusion & Future Work

## 7.1 Conclusion:

The thesis has made use of dataset obtained from Kaggle. The dataset is divided into training and testing phase. 80% of the data is used for training, while 20% is used for testing. After pre-processing and reducing features, a machine learning-based model is built using a convolution neural network. The R2 score obtained is very good as compared to the results obtained from the literature I studied. Hence it can be said that using a convolution neural network we have definitely boosted the accuracy of the model to predict stock price by reducing the error rate r2.

## 7.2 Future Work:

In future, the accuracy can be increased more by using a better hybrid model with CNN algorithm. The model developed can be transformed into the software. Twitter API can be made use like Tweepy for live data analysis so that the model is kept updated with live twitter data. The GUI developed can be improved for better functioning and better user usability.

# References

* Adel, H. & Schutze, H., 2017. Exploring different dimensions of attention for uncertainty detection. *EACL.*
* Anon., 2016. Pew Research Center, News Use Across Social Media Platforms 2016..
* Anon., 2020. NumPy Introduction. Issue https://www.w3schools.com/python/numpy\_intro.asp.
* Anon., n.d. *Python - GUI Programming (Tkinter).* [Online]   
  Available at: https://www.tutorialspoint.com/python/python\_gui\_programming.htm
* Anon., n.d. *Twitter Usage Statistics.* [Online]   
  Available at: https://www.internetlivestats.com/twitter-statistics/#:~:text=Every%20second%2C%20on%20average%2C%20around,200%20billion%20tweets%20per%20year.
* Bowman, S. R., Angeli, G. & Potts, C., 2015. A large annotated corpus for learning natural language inference. *ResearchGate.*
* BRITZ, D., 2015. *Understanding Convolutional Neural Networks for NLP.* [Online]   
  Available at: http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/
* Doshi , Z., Nadkarni , S., Ajmera, K. & Shah , N., 2017. TweetAnalyzer: Twitter Trend Detection and Visualization. *IEEE.*
* Edwards, G., 2018. *Machine Learning | An Introduction.* [Online]   
  Available at: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0
* Fama, E., n.d. The behavior of stock-market prices, The Journal of Business.
* Heller, M., 2019. *What is Keras? The deep neural network API explained.* [Online]   
  Available at: https://www.infoworld.com/article/3336192/what-is-keras-the-deep-neural-network-api-explained.html
* JAIN, K., 2015. *Scikit-learn(sklearn) in Python – the most important Machine Learning tool I learnt last year!.* [Online]   
  Available at: https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/
* Jang, B., Kim, I. & Kim, J. W., 2019. Word2vec convolutional neural networks for classification of news articles and tweets. *IEEE.*
* Lei, Z. & Wang, L., 2015. Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm. *IEEE.*
* Mankar, T., Hotchandani, T., Madhwani, M. & Chidrawar, A., 2018. Stock Market Prediction based on Social Sentiments using Machine Learning. *IEEE.*
* Mittal, A. & Goel, A., 2012. Stock Prediction Using Twitter Sentiment Analysis. *IEEE.*
* Pagolu, V. S., Challa, K. N. R., Panda, G. & Majhi, B., 2016. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. *IEEE.*
* PARKER, T., 2020. *Investopedia.* [Online]   
  Available at: https://www.investopedia.com/financial-edge/0712/10-twitter-feeds-investors-should-follow.aspx
* Roslan, M. A. A. & Rahiman, M. H. F., 2018. Stock Prediction Using Sentiment Analysis in Twitter for Day Trader. *IEEE.*
* Socher, R., Perelygin, A. & Wu, J. Y., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *ResearchGate.*
* Vu, N. T., Adel, H. & Gupta, P., 2016. Combining recurrent and convolutional neural networks for relation classification.. *ResearchGate.*
* Yih, W.-t. & Richardson, M., 2016. The value of semantic parse labeling for knowledge base question answering. *ResearchGate.*
* Yin, W., Kann, K., Yu, M. & Schutze, H., 2017. Comparative Study of CNN and RNN for Natural Language Processing. *IEEE.*