National College of Ireland

Co-relation between the financial news articles and Stock prices and Stock Prediction

MSc Research Project
Data Analytics

Sagar Totannanavar
Student ID: X18104177

School of Computing
National College of Ireland

Supervisor: Prof. Manuel Tova-Izquierdo

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Sagar Totannanavar
Student ID:	X18104177
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Prof. Manuel Tova-Izquierdo
Submission Due Date:	12/08/2019
Project Title:	Co-relation between the financial news articles and Stock prices and Stock Prediction
Word Count:	XXX
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Co-relation between the financial news articles and Stock prices and Stock Prediction

Sagar Totannanavar
X18104177

Abstract

Stock market prediction has been the most shortsighted field. Everyone wants to earn more from their stocks which brings the need for an efficient prediction algorithm which can consider all the possible factors affecting the price of the stocks. It has so many factors which makes it spontaneous and dynamically driven. Till now all the predictions had considered the historical data and the technical indicators of the stock prices and fore-cast the the market and also have used the twitter tweets and the financial news about the companies and analysed the text to predict the trend of the market. This research focuses on combining both that is text analytics and the time series forecasting of the stock prediction and predict the market trend. This research concentrates on implementation of the prediction model using the newly available convolution neural network called the Temporal Convolution network (TCN). This project implements the stock prediction system on TCN whihc is most efficient algorithm compare to the Recurrent Neural Network(RNN), Long Short Term memory neural network(LSTM). Research concentrates on combining the information from the financial news and technical indicators calculated from the historical data of the stock prices. Here we have considered the news about companies over a period of six months and the corresponding stock prices on that day coming days. The model gives an accuracy of 58.3% and also considerable improvement in the time taken and also it is compared with other neural networks.

Area: Text Processing, Semantic Processing, Sentimental Analysis, Prediction, Temporal Convolution network, Event Embedding

1 Introduction

We have known that how crucial the timing of selling and buying of a company's share is. The prediction of the stock market has highest order of necessity in terms of its accuracy and the prediction speed, also the features it considers for prediction. Before the sentiment analysis came ,the talks, press releases of companies, etc were analysed by human and analysed and took action on the stocks depending on their experience in the business, which was not so accurate which was subjected to higher risk. As the technology and science developed there came all the statistical measures and prediction system into existence. Now the technology has grown so much that the prediction do not depend only upon the historical prices it also depends on the mood of the investor after knowing some important news about the respective company.The idea of considering the human

sentiments to predict the stock market is not new, Shiller et al. (1984) has researched that the way investors analyse the stock market includes the social mood around them. And Stock market prediction using the Neural networks was first proposed by Lawrence (1997) where he researched about prediction of stock market using the Neural networks using the historical prices and technical indicators as the features. So as the technology evolved there are many models to predict the stocks. Now in this research we concentrate on extracting the main features from the news and using the most recent architecture of neural network TCN.

1.1 Motivation and Project Background

According to Malkiel (2007) the stock market and stock prices are highly unpredictable. Stock markets are the collective result of all the information available and the changes are very random which is called as the Random walk theory.As the time elapsed researchers have found that stock markets can be predicted. According to the researcher Nofsinger. (2003) the financial decisions of the people depend upon the mood and the emotions they are carrying wit them which is called as the Socio-Economic theory. And the researches also found that although there is evidence that the social mood of the individual plays important role in the investment of stocks in the last decade researchers have kept the social media analysis away and considered only the historical data (Devitt and Ahmad; 2007).According researchers Porshnev et al. (2013) even though historical data and technical data are important for stock prediction, the investor need some more features which include information from the company's decisions and social mood.

According to researcher Das (2010) the financial articles , words used in that article and the emotions , sentiments attached to it makes the difference in behavior of the investors. Including these factors into prediction model will surely increase the success rate. Apart from these researchers there are so many instances which proved that the important news about the companies like buyouts, sales, mergers etc. really changes the condition of the sock of a company drastically. For example in India during 2014, Nestle Limited suffered a big blow, there was inspection conducted on the Maggi which is the product of Nestle India, the report said that it contains some of the banned substances in a higher quantity than allowed and the Supreme court has ordered investigation. This news broke out in all media channels and then it was misinterpreted and this lead to chaos like Maggi is banned in India (Garg; n.d.). This lead to defamation of the company and also investors took out their money out and the market was down to 23% from 63% (Fry; 2016) like this the if we take only the historical data into consideration for the prediction then it may lead difference in many situations. So to include such condition in prediction model we need to include the feature from the news articles into numerical measure before predicting the stocks (Schumaker et al.; 2012).

With the above motivation when searched for the researches done in this area found many related works. In 2010 Bollen et al. (2010) has analysed the mood of the people from the tweets using the GPOMS(Generating Profile of Mood states) and used Granger causality and Neural network and got a higher accuracy than the previous models. Next Xue Zhang (2010) extracted the emotions from the tweets using the words Fear, hope and worry. This model gave better results than the time series predictions. On the concepts these two models in 2013 Porshnev et al. (2013) classified the tweets using Dictionary

model and Naive Bayes theorem. Dictionary model contained the 6 basic emotions. then SVM was used for predictions and it gave far better results the previous models. So for many years many have considered the social media that too twitter extensively but very less on News articles.

But with the advent of technology has changed the tables now. Artificial Intelligence, Machine learning, Deep learning has made the prediction of stock market possible and more accurate than ever. In 2018 Zhang et al. (2018) used the tensor factorisation and matrices coupling to analyse the tweets and news articles and predicted the stocks which gave them a very good result. In the same year Oncharoen and Vateekul (2018) proposed an other model which used event embedding as the feature extraction technique, using CNN and Long short term memory (LSTM) methods and predicted the stocks. So these all gave the idea of combining the features from the tweets and news articles with the numerical data. This project proposes and uses the Temporal Convolution network(TCN) (Bai et al.; 2019) which is new architecture of deep learning. This neural network is the super combination of the best practices of CNN and RNN which takes less memory as compared to LSTM.Also gives the flexible receptive fields which stops the leakage of data from future to past (Bai et al.; 2019). This research paper implements the TCN and uses the news articles from Reddit and stock prices from the Yahoo finance. Also all evaluation measures like accuracy, precision, ROC curve, efficiency are measured.

1.2 Research Question and Research objectives

RQ.1:How the financial news articles tone and subject, mood state of people along with the quantitative historic data, increase the accuracy of stock market prediction.
The financial articles of the companies are taken and the event embedding is used to extract the features from the news articles and then fed to TCN to classify them.

SubRQ:How Temporal Convolution neural networks based models can be used to increase the efficiency of the stock market prediction using the news articles along with the historical data. The technical indicators are calculated using the historical data from the yahoo finance and then combination of both text data and technical indicators are fed to build the model.

1.3 Research Objectives

1.News articles are taken from the financial news websites and are pre-processed to the event embedding vectors.
2(a).The historical stock prices are taken from the Yahoo finance and technical indicators are calculated.
2(b). Event embedding vectors and the technical indicators are fed to Temporal convolution network to train and build the model.
2(c). Test data is given to the TCN model and evaluated fro the prediction values.

The research paper is structured into following sections, Literature Survey which include the critical review of the relevant literature available, Methodology describes the KDD (Knowledge discovery) process,Design Specifications defines the specification of the model, Implementation gives the detail about libraries and process used , Results evaluated in the Evaluation section and lastly the conclusion and future work is discussed.

2 Literature Review on the stock market prediction using the historical data and the social media and news articles (1984-2018)

2.1 Introduction

To start with the related work this paper will discuss about the different works and methodologies proposed by all the other researchers in the last few years.The critical review is divided into following sub sections 1.Literature review on the evolution of stock market and the prediction. 2.Literature review on feature extraction from social media, news articles and stock prediction. 3. Stock market prediction using the deep neural network and text analytics. 4.Conclusion

2.2 Literature review on the evolution of stock market prediction and the social mood:

According the research done by Shiller et al. (1984) most of the investors are not statistically or mathematically qualified. The analysis and the understanding of their stock market movement depends on many factors. In those factors the news articles and the social media news about any company is one of the important factor. Also author justifies that there is no other way in which way the stock markets can be evaluated so the social mood around the investors big impact. So the sentiment of the people towards news articles make impact on stock market prediction.

Adding to the research above, the paper "Social Mood and Financial Economics" it is evident that one of the main factors which plays important role in making decision about buying and selling of the stocks is the emotions of a person and the social mood around him while he is investing. The more positive the emotions surrounding him the more optimistic investor will be and this result in buying any stock (Nofsinger.; 2003). So from this theory it is deduced that social mood around him and the news articles and social mood plays important role in stock market.

Researcher Malkiel (2007) in the paper "A random walk down Wall street" says that the market is very unpredictable and difficult to beat it. It has many factors which affects it. So stock market is very dynamic and changes spontaneously. It difficult beat it but it can be analysed. But there are many examples of different models which are almost close to the results. This proves that the stock market prediction is not only depending on the historical data but also on the other factors like the political policies, decisions taken by the company and the people's reaction towards it. Lets see the way prediction evolved in the next section.

2.3 Literature review on feature extraction from social media, news articles and stock prediction:

As the machine learning and deep learning are invented researchers started using them to predict stock markets. In 1997 Lawrence (1997) proposed a prediction model using

neural network to predict the stock market. Author claims that the Efficient Market Hypothesis(EMH) assumes the fully granulated random information which leads to the unpredictable market. But in reality there is a huge amount of data compared all other theoretical information using which one can build prediction model.Author built the different neural network recurrent network, generative network, hybrid network etc. The result of these network was calculated using the SP500 Index, for which the result increased by 10%, and network with back propagation and recurrent networks returned 30%.

In the paper, Using News articles to predict stock price movements (Gidfalvi; 2001), author has proposed method which is totally different from the usual way. Here he used the news articles or the text as the independent variables and then applied the Naive Bayes classification on them. This paper they derived 3 identifiers. they are aligning , scoring and labelling the articles. And at the end model is trained using Naive Bayesian classifier for the movement of classes and the posterior probabilities are calculated and used. Overall scoring included the volatility of the stocks, which was beta value. News articles are labelled by seeing the change in the stock price along with the index price change. Using the Baysian model they were able to get the better results for the first twenty minutes after that the stocks were remaining same. Still this was a god model where for the first twenty after market has started there was prominent change and the users can be benefited from that.

Next the paper gives the idea about extracting the features from the social media articles and using them to classify the tweets, news articles and predict the stock market trends. In 2010 researcher Bollen et al. (2010) thought to use the mood states from the tweets and then use them to build the model and classify the tweets and predict the stock market trend. Here the author has used the Opinion finder tool and Google - Profile of Mood states(GPOMS) to extract the mood of the people in social mood regarding news of a company. Here the data set used to experiment was of Dow Jones Industrial Average(DJIA) over a time, data set contained the tweets from pope regarding the news related to DJIA and the stock rices of DJIA at that time. From the opinion finder tool the author was able to get the positive vs negative time series of the public mood, then they used the GPOMS which analysed the tweets and gave a 6 dimensional array which contained the co-efficient for calmness, alert, sureness, kindness, vital and happiness. This gave a detailed view of the public's mood compared to the opinion finder. Then these moods co-efficient were correlated against the stock price time series to get some relation and build model. The results seen were as following: adding new mood dimension increased the correlation between moods and the stock price, while for opinion finder it was not efficient. Even though this method gave a 6 dimensional array of moods but there was no proper method extracting the moods from the tweets. So this method can be considered as the basic method.

After the first attempt of extracting the features from the tweets the text analytics grew exponential and many new type of methods came into existence. In that let's consider researchers Xue Zhang (2010) in their work "Predicting stock market indicators through twitter I hope it's not as bad as I fear" used the social media twitter as the source of mood and news about the DJIA, NASDAQ and SP500. And they also considered the stock prices for the same. Tweets were collected for the period of 6 months and then it

is evaluated using the words Fear, hope and worry as the measures. A per the author E and K (2010) the emotional state of an individual will influence the decisions taken by him. When ever the people are pessimistic or worried they become cautious and think many times before they invest. So considering these factors is very important and it can help in predicting the stock markets. On this back ground the researcher Xue Zhang (2010) gave importance for the words far, worry and hope, and counted the number of mentioned words in the tweets in a day are calculated. and they were classified then the stock value for the same day is considered and compared , they found that when ever there are words like fear, hope and worry the stock price fluctuates negatively, which says the mood and the stock price are related. The lag in this research was the number of samples considered. For retweets the number was less and it lead to the difference in results.

Next in the year 2012 the researcher Schumaker et al. (2012) proposed a model called AZFINText model. This model was built using the SVR sequential minimal optimisation function through Weka. In this proposed model researchers used the stock price data and the news related to them from a commercial website. The news articles are analysed to get the sentiments and other pronouns. Here they used the opinion finder tool to analyse the text's nature that is a sentence is subjective or objective . Then again the same tool is used on the analysed text and the polarity of those subjective or objective texts is found out. And then this data is given to the time series module Weka. The result from this model was discrete and it also predicted the value of share for next 20 minutes. The results founded were subjective news had more impact on the investors and it impacted the stocks. Also the negative news made more impact on the stock market. Some time the news articles had the contrarian effect also, that is even though the polarity was negative the result , action of the investor was positive. The overall take away was subjectivity of a text matter more than its objective part. Some improvements needed for this model was in taking the verbs into consideration while analysing texts.

2.4 Stock market prediction using both historical data and text analytics:

As the technology evolved researchers found out the importance of both the historical data and the textual data in the Stock market prediction. In that direction the researcher Wang et al. (2012) proposed a novel model which used both linear and non linear data in constructing the model. In the paper "A novel text mining approach to financial time series forecasting" author used the dataset of 6 companies which is texts release and the ROE numbers.In this model author considered the output of the model as the combination of linear(time forecasting of stock prices) and non linear part(text analysis of the news articles). Linear part was done by the ARIMA implementation and the non linear part by the SVR(Support Vector Regression model) the result from this model was higher than the single ARIMA model and also text analyses alone. The drawbacks were it was suited for only smaller datsets. And also there will be some data loss when we try to select the relevant data which is corresponding to time series data. From this the era of combining the text and numerical data started.

On the Schmakers's model AZFinText model (Schumaker et al.; 2012) Proshnev and

others (Porshnev et al.; 2013) proposed their new model which used the WFH words and also the 8 emotions Lexicon to analyse the text or news articles and tweets. After that used the SVM and Neural networks to predict the trend of Stock. Here the authors have consumed the same library used by Bollen et al. (2010) to analyse the tweets and sentiments.Here they combined three methods of text mining, they are WFH emotions, 8 emotions from lexicon and Dictionary method for classification. Then used the SVM to predict and got better results than Bollen et al. (2010). even though the overall results were better than all others it has a place for improvement in consideration of the relevant emotions.

In 2014 researchers Ding, Zhang, Liu and Duan (2014) proposed a new way to predict the stock market using both the news headlines and the historical data. In this paper researcher used the Google's Open Information extraction method to extract the events or features from the collected data. Researcher Kim in 1993 has proposed this method of getting the structured events, this has been used by the authors. The event extraction helps in getting the syntactical and semantic features from the news. In this method the sentence is analysed to get the action, verb and actor. To analyse the Non liner part of the mode they have used the deep neural network and for the linear part have used SVM for the prediction model. The overall experiment of this research shows that event based analysis of the document is the efficient way (Ding, Zhang, Liu and Duan; 2014).

In 2015 the researchers Ding et al. (2015) proposed a new model where they used CNN as the neural network and used the Event embedding technique for feature extraction from the textual information. In the previous works the event structures which were extracted from the texts were very sparse to improve that here hey have proposed the event embedding technique which gave a dense embedding vectors. Here the sentences are broken down into 3 parts action, actor and object. From these the action and the object are fed to the model so that embedding can be formed. These embeddings then fed to the layer of CNN for training. Each layer in CNN is event embedding layer , these taken the input and give a U for that day. To correlate the features vector and stock prices a feed forward neural network with hidden layer and 1 output layer. Vc = (V1, V2, V3) are the averaged feature vectors and $Y = \sigma(W_t, V_c)m$ is sigmoid weight vector. From this model authors got very good results.

AS development evolved the scientists Matsubara et al. (2017) developed a LSTM network. In this paper they have discussed the paragraph vectoring which means converting the paragraphs into vectors so that whole context/semantics can be captured efficiently. And one more short coming from the other models was SVM, SVM cannot take the timeseries features into consideration. So here SVM is replaced by the deep neural network. The paragraph vectors and the stock price indicators are both fed to the neural network and combination of these two gave a better result than all other models.Also in this paper authors considered the prices for different companies instead taking the values from single company. (Matsubara et al.; 2017)

In 2017 Li et al. (2017) proposed a technique called SMeDA-SA (Social Media Data Analysis - Sentiment analysis) to analyse the people's sentiment from the social media data.Here they considered the data from the twitter to extract the textual features like the mood of the people after listening to any news. Here they considered 30 companies

which are listed publicly, and stock prices of those companies also collected. Here the collected data is divided into five categories and the ambiguity is removed and data is analysed efficiently. the main of the research was to remove the ambiguity of the social media data and the sentiment of the people with the companies are calculated and this data was correlated with the stock prices of corresponding company. and it predicted the stock prices. This model gave better results but as per the news articles and the tweets it is better to consider the semantic information also and there is a need to consider how the events are affecting the stocks. So in that way the following model given by Oncharoen and Vateekul (2018) feels better.

In 2018 researchers Oncharoen and Vateekul (2018) proposed a new model which was the combination of both the textual information and the stock prices. Here they used the LSTM to build the Model. In this paper they have used the event embedding technique to get the features from the news headlines. The events are taken into consideration and embedded with stock prices to predict the stock values. Here they have used CNN and LSTM layer separately. CNN was used for the event embedding and LSTM for the stock price prediction.It gave better results as compared to the models which used the historical data alone.

Now after learning that news articles and twitter news financial news are as much important as the historical data or the stock price, now it is time to learn which neural network is better to calculate the stock prices and the events relation. So in that direction this research project considered TCN as the better one. In the following research paper 'An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modelling' authored by Bai et al. (2019) have proposed a new architecture called Temporal convolutional network, which can solve the sequential task with better efficiency than the LSTM in terms of length of retention , speed and the memory. The proposed network has outperformed the RNN(Recurrent neural network) in the initial point high power. This architecture has the best practices from CNN and RNN. Following are the main features of proposed TCN 1. The convolutions in the architecture are causal, which means that the there is no leakage of data from future to past or vice versa. 2. It can take input of any length and process it as single output.
Convolutional networks (LeCun et al., 1989) were used to analyse the sequential tasks for decades (Sejnowski Rosenberg, 1987; Hinton, 1989). In 80s, 90s the convolution networks were used for speech recognisition systems (Waibel et al., 1989; Bottou et al., 1990).Convolutions Network were sequentially applied to NLP tasks such as part-of-speech tagging and semantic role labelling. This architecture is simpler than the wavenet (van den Oord et al.,2016).
The architecture is built on following two factors 1. Sequential modelling which means the predictions of event can be done on the past events. 2. Causal convolutions:It has 1D causal convolution. The memory of the network is increased by this feature. It can accommodate more filters which means more number of past events. Next it has dilation layer. Dilated convolutions help in having larger receptive field which also help in increasing the memory so that the model can look back into the past events and learn the pattern. TCN has the flexible residual connections.These help in learning the modifications to map the changes than entire entire transformation. TCN has advantage having parallelism, flexible receptive field size,stable gradient size, low memory for training, variable length inputs.

Two disadvantages are storage during evaluation, potential parameter change for a transfer of domain. after considering all the things we for with the TCN, As here the history needed is more and the time to be taken for calculation and the memory should be less. (Bai et al.; 2019)

2.5 Conclusion:

From the review of the related work it is found that as the technology developed the methods used for the feature extraction also improved. The text analytics got powerful with all the techniques and it helped in the extraction of exact features. Also the prediction of stock market also improved with the different models.

First the tweets were analysed and emotions were collected and predictions were done (Shiller et al.; 1984). Then came the method of Lawrence (1997) who did the time series forecasting of stock market. Initially Porshnev et al. (2013) used the machine learning techniques to analyse the data and text and predicted the stock market. Eventually text analytics improved and many forms it came into existence they word embedding Ding, Zhang, Liu and Duan (2014), using the deep learning techniques the efficiency of stock prediction increased. In 2014 researchers Ding, Duan, Liu and Zhang (2014) used Open IE for feature extraction. And then era of deep learning started which increased the efficiency by the LSTM, CNN together geve better result for Oncharoen and Vateekul Oncharoen and Vateekul (2018). After comparing all the things we came to TCN which is the new architecture which was implemented by using the best qualities of CNN and RNN (Bai et al.; 2019), this gave the motivation to build a better prediction system which takes the human emotions into consideration and also the technical indicators.

3 Methodology

To implement the data mining project we follow the Knowledge discovery in database (KDD) process. It is the combination of all the steps which are needed to extract the knowledge from a database or data sources. The KDD process and the steps followed in it are explained by (Fayyad et al.; 1996). This section describes different scientific methods followed in this research to get the knowledge from target data sources.

3.1 Data Collection:

For this research , target data sources are taken YahooFinance[1] and the news data from github repository[2]. The financial historical data from the Yahoo finance was collected for the period six months.Then the news data from the github which is collected by scraping the websites like reuters, economic news, etc financial news articles datas is used for the text analytics.

[1]https://finance.yahoo.com

[2]/https://github.com/gyanesh-m/Sentiment-analysis-of-financial-news-data

3.2 Pre-processing:

In this the data from both Yahoo finance and the github are pre-processed and made ready to extract the meaningful features from them so that it can be given to the models as input to gather knowledge from it. The steps followed are:

1. Removing all the null values: The unwanted noise which is present in the collected data that is null values are removed. This made the data more readable, meaningful and helped in getting the better predictions accuracy.

2. All the special characters, URL links, unwanted texts in the news headlines, embedded image links, name tags are removed so that the main features can be extracted efficiently and texts can be analysed to get the events related to stock prices.

3. Case conversion of the articles , so that the words in different form can be transformed same form so that all the action words are collected.

4. The words with contraction: The words which are in short form like didn't, ain't, I'm are transformed to their original form so that the meaning and action of the sentence remains same and gives more information and better word tokens when they are transformed.

5. Word Tokenising : The process of word tokenising of news articles is very important task. The news headlines are broken into word tokens so that the features are extracted from each articles properly without loosing the weightage and meaning.

6. Removing the helping verbs, stop words like to, from , and etc.,so that the news articles become very much ready and easy fro the extraction of features.Also retweets and and multi copies of same article are removed.

7. Replacement: Replaced the elongated form of a word with its original form so that the important features are not lost.

8. Stemming and lammetising the words: In this steps all the words are transformed to their root form so that proper meaning and a good dictionary can be built.

3.3 Feature Extraction and Transformation:

Here in this research features from the text headlines using different techniques like word embedding, event-embedding, Open Information extraction methods to get the required action and events from the Texts. In this project features are the events like Microsoft buying any company, Samsung releases new galaxy series, from such news the events are extracted. Which can be transformed and used as input for the model. To extract the events and then vectors from them. Event embedding technique was first proposed by (Ding et al.; 2015). As proposed by Ding et al. (2015) the information extraction is done by using the Open IE process proposed Ding, Zhang, Liu and Duan (2014).Using the open IE we get the words which are then converted to vectors using the word embedding, these vector tuples which contains three objects Actor, Action, Object. Using these event embedding vectors are created .
Transformation: Next is transformation, this is the step where the features are transformed into required form to feed as input to the models.The transformation techniques

are converting the Subject,verb, object into embeddings. Firstly event vectors are created by feeding the Actor (A_1),Action(A_2)and Object(O_1) vectors to a neural network. The Neural vector will produce the vectors (R_1,R_2) by combining the Actor, Action and object vectors and multiplying the Action and Object by a tanh function produces a vector of actions. These actions vectors is again multiplied by tanh function considering the weight matrix and nullifying the and final Vector of events is given.The vector R_1 gives the relation between Actor (0_1) and Action (P). This vector is computed by equation (1) where f is tanh function, k is a dimension of the input vector, W is k * 2k weights matrix, and b is biases vector. Like this the vectors R_1, R_2 vectors fed to the neural network which will convert them to a single event vector called U, Equation for event vector U describing the relation between $R_1 and R_2$ is given in Equation(2) (Ding, Duan, Liu and Zhang; 2014)

$$R_1 = f(O_1^T * T_1^{[1:4]} * P + W[{}^{O1}_{P}] + b) \tag{1}$$

$$U = f(R_1^T * T_3^{[1:k]} * R_2 + W[{}^{R_1}_{R_3}] + b) \tag{2}$$

Next part is converting the technical indicators and the historical prices into a vector. This project has considered the technical indicator proposed by Vargas et al. (2017). The below image shows the table indicators used. These formulae and the indicators are taken from the (Oncharoen and Vateekul; 2018). Project have considered the normalised value of the technical indicators and the prices using the z scores. As the price and the indicators are in different range it is necessary to normalise them before using them in the equations.Figure 2 shows the conversion of the technical indicators.

Feature	Formula	Feature	Formula
Stochastic %K	$\frac{C_t - LL_n}{HH_n - LL_n}$	William's %R	$\frac{H_n - C_t}{H_n - L_n} x100$
Stochastic %D	$\frac{\sum_{i=0}^{n-1} \%K_{t-i}}{n}$	A/D Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t}$
Momentum	$C_t - C_{t-n}$	Disparity 5	$\frac{C_t}{MA_5} x100$
Rate of Change	$\frac{C_t}{C_{t-n}} x100$		

where C_t is the closing price at day t, H_t is the highest price at day t, L_t is the lowest price at day t, MA_n is moving average of the past n days, and HH_n and LL_n are the highest high and the lowest low in the past n days, respectively.

Figure 1: The conversion of the indicators (Oncharoen and Vateekul; 2018)

3.4 Data mining:

In this step the data which is preprocessed and transformed is fed to the TCN layers for training and creating the model and testing it.This research employed the TCN model which takes the features from the textual information that is infromation from the news headlines which are converted into event embeddings from the Subject, object verb phrases (Ding et al.; 2015).And the historical data which is converted into technical indicators. The event embeddings along with the technical indicators are fed to the TCN. The flow chart of the model is shown in the Figure 1.

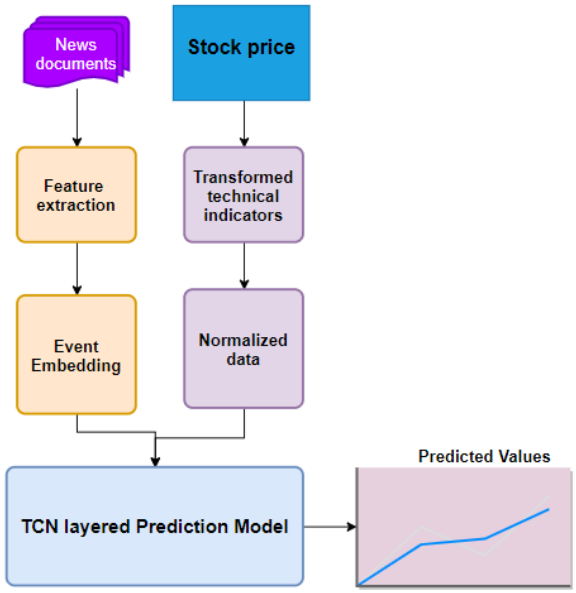

Figure 2: Design of the proposed model

The main deep learning neural network considered to build the model is TCN(Temporal COnvolutional Network). TCN is the architecture which is made by selecting the best practices of CNN (Bai et al.; 2019). At first recurrent networks and its varieties were considered to be best suited for the sequential modelling but as studies were done, TCN are seem to beat the RNN and its varieties in sequential modelling (Bai et al.; 2019). The Main characteristics of the TCN (Bai et al.; 2019) are :

- The convolutions are leak proof or causal. That is there is no leakage of data from the future to past or vice versa.
- TCN netwrok can handle input of any length and can also build the sequential output of same length, as the it a large history.

Advantages of the TCN over RNN and its variants:

- Parallelism : In TCN the predictions for different convolutions can be processed at the same time parallely with other convolutions as a whole.
- TCN has the flexible receptive field size, which means it can contain many convolutional layers and also the filters. Filters help in increasing the memory size of the model
- Gradients: The back propogation way of the sequence is different in TCN as compared to RNNs. The propogation does not take the direction of temporals . It helps in speeding up the evaluation.
- Memory requirement for training: These consume less memory for training as compared to RNNs as these have shared filter and also the path of back propagation is also dependent on the depth of the network.

- Length of the input: These can take the inputs of any length by sliding the 1D convolutional kernels. Hence can be used as the replacement for RNNs. With all these features and the advantages this project has chosen the TCN technique to predict the Stock prices.

As per the below model the transformed data from text and the historical data that is fed to the TCN layers The transformed event embeded vectors are fed to the first layer TCN and output from this are fed to next layer, on the other part the technical indicators are fed to the LSTM network for tim series prediction. The output from these are fed to the last hidden layer of TCN which concatenates with the outputs from the event embedding vectors , Then the model trained such that the predictions whose values are more than one are classified as class+1 and whose values which are less than 1 are classified as class-1. This is the proposed approach. And prediction values are evaluated for accuracy.Diagram in the Figure.3 shows the structure of the model.

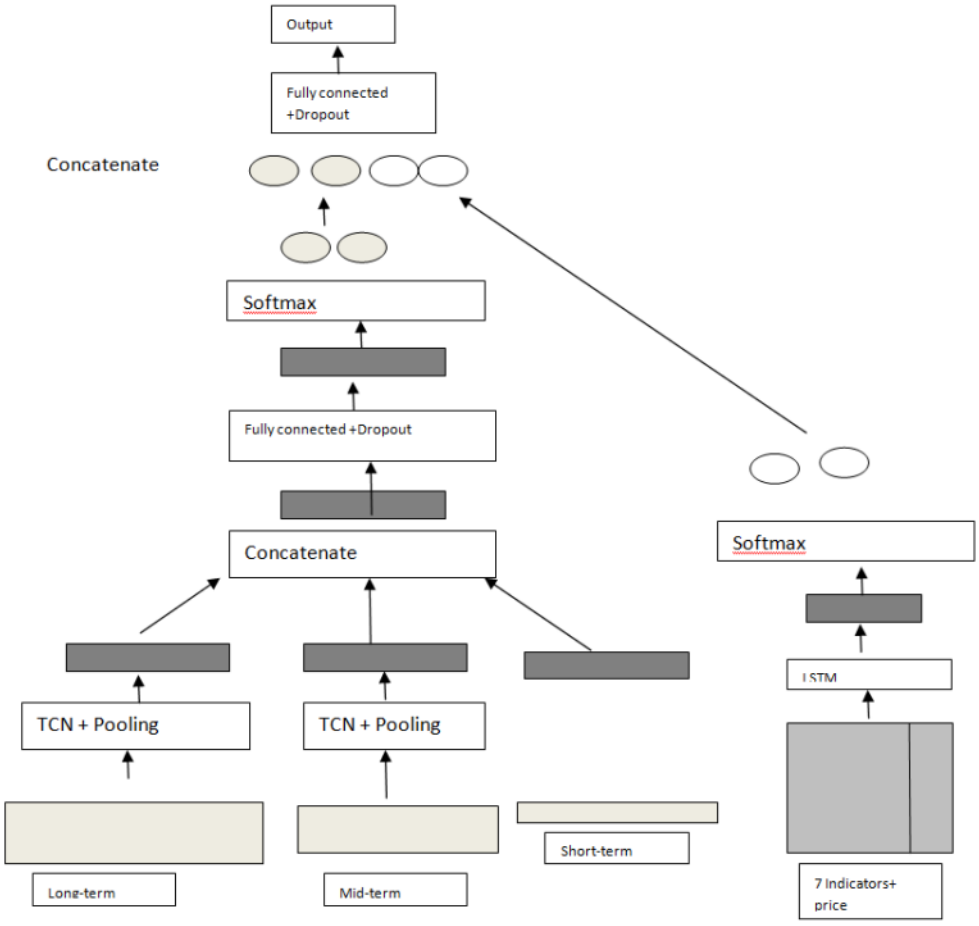

Figure 3: Flow chart of the model

4 Implementation

4.1 Introduction:

In Implementation section all the details about the research implementation is discussed. This section has subsections like Dataset: in which the details about data set like source of

the datset, duration and the information what is gives is described. Data Preprocessing: In this section all the libraries and the procedures for preprocessing the data are explained. Event embedding: In this section the implementation of the feature extraction from the news headline is explained.TCN Model: In this section the libraries, methods, steps and the languages used for building the model are explained.

4.2 Data Set:

1. The historical data is taken from the free stock data provider that is Yahoo Finance! website. The collected data is the Stock indices of Nifty50 , Nifty50 midcaps, Nifty small caps . This is the stock market indices of Nifty which is group of the companies which is based on market capitalisation of 10 large companies.This historical data is collected for a period of 1year.
2.For News headlines : Financial news headlines from Reuters. Have to select only the headlines that contain company names or abbreviations.
3.The News dataset from the github is taken which is collection of all the news related to the top 10 large companies of NIFTY50.

4.3 Data Pre-processing:

Preprocessing of the raw data collected is very important. As it contains much noise the data should be cleaned and transformed into a proper form which is usable in the models for analysis. To preprocess the data, have used both Python and R programming languages. At first the null values from the financial news dataset and stock price data is removed.The stock prices and financial news are joined together by date using the programming language Python.
Then the special characters, numbers, URLS, unwanted tags are removed from the news headlines using the tmmap, tidyverse, qdap libraries in R language. punctuations, are removed and then the stop words are removed from the news headlines.
Next the stock prices are taken unwanted values are left. The technical indicators are calculated using the methods available in the R programming language.

4.4 Event Embedding:

This is the main step while extracting the features from the textual data. To get the event embeddings we followed the following procedures : The event embedding is implemented using the python programming. Have Used the libraries Pandas, Numpy libraries to access the data and convert it to arrays.
Library textacy, spacy are used to to convert the sentences into subject, verb and object form so that the word related to these part of speech are collected. pickle library is used to save the models of SVO tags. These are saved.
Next the word vectors are generated for the svo tags. This is done using the google Glove model. which is vocabulary model formed by the google. The transfer learning is used and vectors for the concerned word is taken from this model. Libraries used are gensim package in that word2vec model is used.
Then these are grouped into three categories like long term, short term, medium termed ones because the data collected was for the midcap, small cap and the longer term one and all the news which are on same date are merged and the k mean of their vectors is

taken.
The Word vector from these are fed to the neural network which is a TCN to train the model. Tensorflow is used as the backend. Tensorflow is useful python package and keras is used to build the neural networks. All the convolution layers are implemented by keras.

4.5 TCN Model:

Here we will be using the Python libraries like Tensorflow, Keras, Keras-TCN for the implementation of TCN.
Tensorflow is the python library which is highly used for neural network to define the high dimensional arrays. Tensor is the library to handle multidimensional arrays. Tensorflow has many other useful functions which will help in making the machine learning model. Keras is used to build model. Keras layers and keras model functions are taken.
TCN model is implemented which has 1D fully convolution network, Dense layers, Dialtions up to 2^9 , kernel size = 2, residual blocks , activation functions and loss functions of choice. And number of epochs also.The loss function used is the Binary cross entropy, activation used is different for different layer, for the 1D convloutional layer 'Linear' is used, for the dens layer softmax is used.

5 Evaluation

5.1 Introduction:

Here in this section the results of the research are explained. We have three steps in our model. One is converting the words into event embedding vectors. It is in itself is model, so the accuracy of that model is discussed.

5.2 Evaluation of the event embedding model:

This is a TCN model which takes the Long term, short and mod term word vectors which are then converted into event embedding vectors. In this we split the dataset into test and train in the ratio of 80:20. Then given to the model.We got an accuracy of 58% for the long term vectors. 60% for the mid term and 62% for the short term vectors.

5.3 Evaluation of the TCN Model

In this the TCN model created is evaluated.The TCN model is giving an accuracy of 58.9%. Wen the dataset of length was for 6 months.Below are the model efficiency parameters

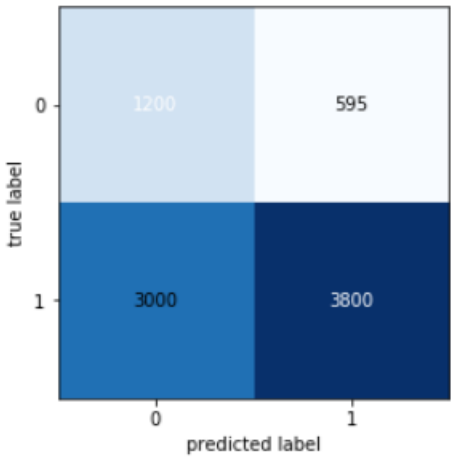

Figure 4: Confusion matrix

- Precision = ability of a classification model to return only relevant instances=Number of true positive/Number false positive +number of true positive
  The calculated precision is 86.4 %
- Accuracy = Number of Correct predictions /total number of predictions = 58.9%
- Receiver Operating Characteristics (ROC) = Number of true positive/(number of true positive+ number of true Negative) = 79.1%
- Recall = ability of a classification model to identify all relevant instances =True positive/(true positive + false negative) = 55.8%
- The F1 score is the harmonic mean of precision and recall taking both metrics into account in the following equation: F1 = 2*(precision*recall)/(precision+ recall) = 70%

Above evaluation metrics tell the efficiency of the model. As we can see the accuracy is less . But when it compared to the models which were considered in (Ding et al.; 2015) paper it is same. Still we have a lot of chance to increase the efficiency.The accuracy here is less because of the dataset length. Here the time consider for forecasting is very less. So the prediction values also varies.
Also the model is evaluated for different hyper parameters like number of epochs, loss function used, activation function used. In this project the model was run for epoch of 1000 , as the epoch increased the accuracy increased from 54% to 58%. Also the activation function was changed and checked for sigmoid and softmax there was very difference in performance. Then it was evaluated with K-fold cross validation. The 10 fold validation was done and the accuracy almost remained same for all the steps. ROC curve was plotted for the model, AUC is 0.65 it is bi less. ROC curve is shown in below figure

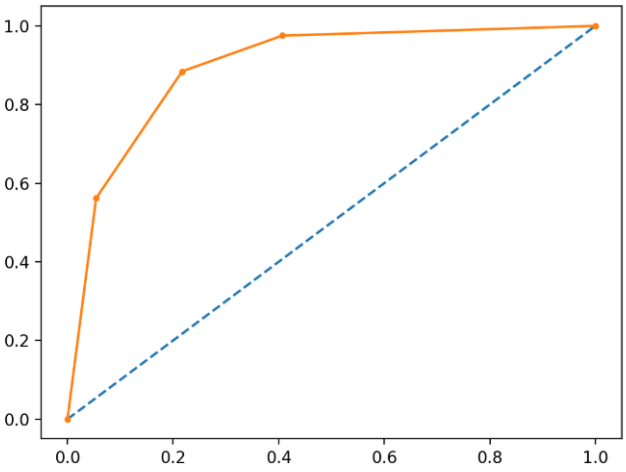

Figure 5: ROC Curve, AUC = .65

The LSTM time series forecasting of the technical indicators which was used for the TCN model. The prediction and test values are plotted as below.It shows the LSTM model which is build for the forecasting is efficient. Its accuracy is 85%.

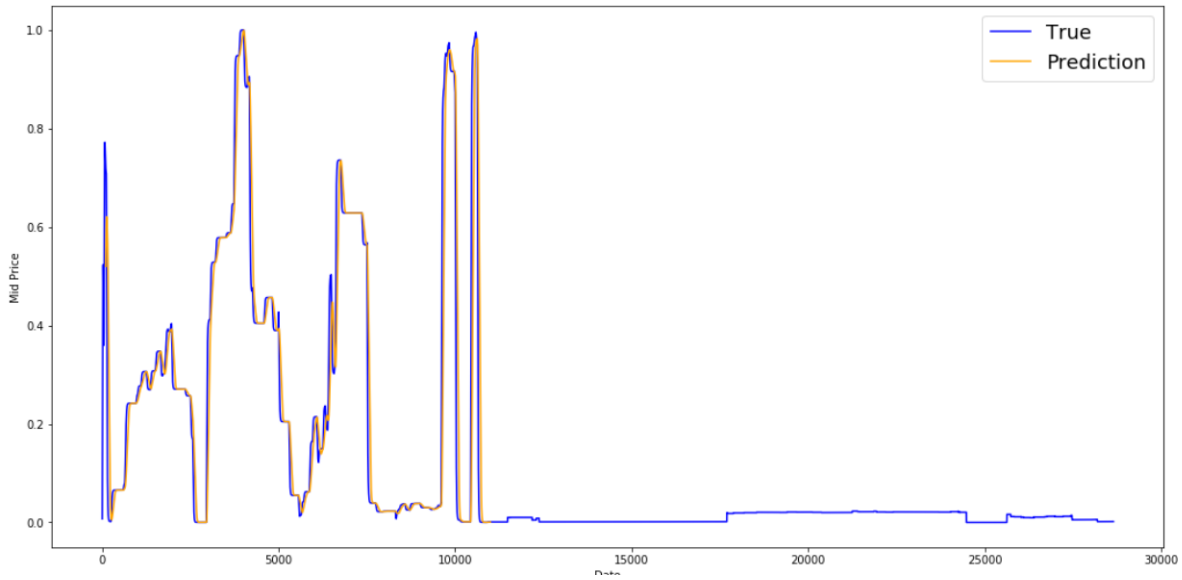

Figure 6: Time series analysis of the data

5.4 Discussion

In the research conducted to analyse the effect of the news articles in prediction of stock market along with the statistical indicators, we have the accuracy of 59%. On the above it seems very less but we compare the results with the amount of data it is built makes the point that there is impact of the news articles on the prediction. When the stock prediction is done only with stock indicators the accuracy was 55% so it was mere a chance. Also the results of the model we have built to compare with this is based on the (Ding et al.; 2015) here in this they got a accuracy of 54% for the short term stocks ,

where they considered only the news, in this model the accuracy is increased by 4%.
To increase the efficiency of the model the news data which is used for analysis can be pre-processed more with the open IE and using the large trained google word vectors to extract exact events which can give better relation with the stock price. In the 2018 Oncharoen and Vateekul (2018) there proposed model with CNN and LSTM gave a accuracy of 63% which was built using the large dataset of DJIA. When we compare the models the efficiency is less, but with the better processors and the large dataset the TCN based event driven stock predictor along with technical indicators will give a very high accuracy.
The main constraints for the research were the hardware , if the GPU processors are used and the number of epochs in both the neural network that is for model building event binding vectors and for the TCN. In TCN the important feature is the number of the blocks of 1D FCN used, and the number of such blocks used (Bai et al.; 2019). By considering more number of convolutions and filters the results will in increasing the efficiency. One more factor to be considered is the time taken for the convolutions and the memory. Here in the TCN the time taken for each convolution is less compared to other models, Which is a main thing in stock prediction.

6 Conclusion and Future Work

6.1 Conclusion:

This research paper has proposed an efficient Stock prediction model which is built on both the technical indicators and the historical indicators/ technical indicators.In this build the news headlines for the Nifty stocks was considered and the events from those are extracted using the Open information and event embedding methods. These events are considered instead of the word vector because the events which all are having the same effect on stocks can be clubbed even though the words are different. This gives the edge over all other methods used for the text analytics. Next is the technical indicators are combined with the news headlines. Both of these are fed to the neural network which was built by the new type of architecture which is combination of RNN and CNN, by taking best of the CNN and RNN called Temporal Neural Network (Bai et al.; 2019). This is family of neural network which has proven a better performance than the RNN while doing sequential tasks. The model built by this architecture has produced a accuracy of 58%.This model gives the best results, as compared to the other models. TCN has the advantages like parallelism, longer history, and can produce very long sequential data. As it has causal convolutions it does not leak any data from future to past.With all these advantages the TCN become the ideal neural network for the sequential tasks. The combination of news articles and the stock price data to predict the stock built by using TCN would be a great help to investors, so that can infer and earn from their investments.

6.2 Future Work:

However the efficiency in this research is almost as compared to (Oncharoen and Vateekul; 2018).But it is because of the length of data used and the platform used for prediction. With the increase in number of convolution layers, filters and other hyper parameters like activation function, epochs the accuracy may increase.

Also the other factor which can increase the accuracy is the text analytics techniques. The sentences can be analysed efficiently using the attention mechanism in text analytics. And also dividing the news events into proper intervals and analysing may increase the efficiency.

7 Acknowledgment:

I sincerely thank my Supervisors Professor.Manuel Tova-Izquierdo and Dr,Sachin Sharma for their support and guidance.

References

Bai, S., Kolter, J. and Koltun, V. (2019). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling., *Cornell University.* .

Bollen, J., Mao, H. and Zeng, X. (2010). Twitter mood predicts the stock market, *Journal of Computational Science* .

Das, S. (2010). The finance web: internet information and markets, *IEEE Intelligent Systems* .

Devitt, A. and Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach, *ACL* .

Ding, X., Duan, J., Liu, T. and Zhang, Y. (2014). Using news articles to predict stock price movements., *Conference on Empirical Methods in Natural Language Processing* pp. 1415–1425.

Ding, X., Zhang, Y., Liu, T. and Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 1415–1425.

Ding, X., Zhang, Y., Liu, T. and Duan, J. (2015). Deep learning for event-driven stock prediction, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 1415–1425.

E, G. and K, K. (2010). Gilbert, e. karahalios, k. (2010). widespread worry and the stock market. 4th international aaai conference on weblogs and social media (icwsm), 2010., *International AAAI Conference on Weblogs and Social Media* .

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data, *Communications of the ACM* **39**(11): 27–34.

Fry, E. (2016). Nestls half-billion-dollar noodle debacle in india.
URL: *http://fortune.com/nestle-maggi-noodle-crisis/*

Garg, N. (n.d.). Impact of maggi row in india.

Gidfalvi, G. (2001). Using news articles to predict stock price movements, *University of California, San Diego* .

Lawrence, R. (1997). Using neural networks to forecast stock market prices., *Department of Computer Science University of Manitoba.* .

Li, B., Chan, K. C., Ou, C. and Ruifeng, S. (2017). Discovering public sentiment in social media for predicting stock movement of publicly listed companies, *Information Systems* **69**: 81–92.

Malkiel, B. (2007). A random walk down wall street.

Matsubara, T., Uehara, k., Akita, R. and Yoshihara, A. (2017). Deep learning for stock prediction using numerical and textual information, *IEEE* .

Nofsinger., J. R. (2003). Social mood and financial economics, *The Journal of Behavioral Finance* .

Oncharoen, P. and Vateekul, P. (2018). Deep learning for stock market prediction using event embedding and technical indicators.

Porshnev, A., Redkin, I. and Shevchenko, A. (2013). Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis, *2013 IEEE 13th International Conference on Data Mining Workshops* pp. 440–444.

Schumaker, R. P., Zhang, Y., Huang, C. and Chen, H. (2012). Evaluating sentiment in financial news articles, *Decision Support Systems* **53**: 458–464.

Shiller, R., Fischer, S. and Friedman, B. (1984). Stock prices and social dynamics., *Brookings Papers on Economic Activity,* pp. 457–465.

Vargas, M. R., de Lima, B. S. L. P. and Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles, *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pp. 60–65.

Wang, B., Huang, H. and Wang, X. (2012). A novel text mining approach to financial time series forecasting, *Neurocomputing* **83**: 136–145.

Xue Zhang, Hauke Fuehres, P. A. G. (2010). Predicting stock market indicators through twitter i hope it is not as bad as i fear, *Procedia - Social and Behavioral Sciences* .

Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B. and Yu, P. S. (2018). Improving stock market prediction via heterogeneous information fusion, *Knowl.-Based Syst.* **143**: 236–247.