# Stock market prediction

**Ravi pandey**

**Computer science**

# ABSTRACT

Several researcher's equally in world and industries have long been interesteds in the stock market. Many approache were developed to accurately predicts future trend in stock price. In recent times, there has been a growing interest's in utilizing graphs-structured data in computer science research community. Method that use social data for stock market predictions have been just proposed, but they are still in their Infancies. First the quality of collected informations from differents type of relation can vary considerably. No current works has focused on the effect of using differents type of relation on stock market predictions or finding an effectives way to selectively aggregate informations on different relation type. Furthermore existing work have focused on only individual stock predictions which is similar to the node classification tasks. To address this we propose  hierarchical attention network for stock prediction (HATS) which use relational data for stock market prediction. Our HATS methods selectively aggregate informations on different relations type and add the informations to the representation of each companies. Specifically node representation are initialized with feature extracted from a features extraction module. HATS is used as a relationals modeling module with initialized node representation. Then node representation with the added informations are fed into a task-specific layer. Our method is used for predicting not only individuals stock prices but also market index movement, which is similar to the graph classification task. The experimental result show that performance can change depending on the relationals data used. HATS which can automatically select informations outperformed all the existing methods

# Introduction

Stock market are  symbol of market capitalisms and billion of share of stocks are traded every day. In 2018, stock worth more than 65 trillion U.S. dollars were traded worldwide and market capitalization of domestic companies listed in the U.S. exceeds the country's GDP 1 . Although stocks movement predictions is a difficult problem its solutions can be applied to industries. Various researcher in both industry and world have long shown interests in predictings future trend in the stock markets. Researcher's focused on finding profitables pattern in historicals data are known as quant in the financial industry and referred to as data scientist in general. Regardless of which terms is used such researcher are increasingly using more systematics tradings algorithm to automatically make trading decision. Even though there is still rooms for debates [21] several studie have showed that the stock markets is predictables to some extents [5], [23]. Existing method are based on the idea of fundamentalist or technician, both of whom have differents perspective on the markets. Fundamentalist believe that the price of securitie of a companies correspond to the intrinsic value of the company or entities [8].

If the current price of a company's stock is lower than its intrinsic value, investors should buy the stock as its price will go up and eventually be the same as its fundamental value. The fundamentals analysis of a companies involves in-depth analysis of its performances and profitabilities. The intrinsic values of the companies is based on its products, sales, employees, infrastructures, and profitability of its investment[2]. Technician on the other hands do not consider real world event's when predictings future trend in the stock markets. For technician stock price are considered as only typical time serie data with complex pattern. With appropriate preprocessing and modeling, patterns can be analyzed, from which profitable patterns may be extracted. The informations used for technical analysis consist of mainly closing price return, and volume. The movements of stock price is known to be stochastic and non-linear. Technical analysis studies focus on reducing stochasticity and capturing consistent patterns can be analyzed, from which profitable patterns may be extracted. Earlier studies on stock market predictions are based on the historical stock prices. Later studie have debunked the approach of predictings stock market movements using historical price. Stock market prices are largely fluctuating. The efficient markets hypothesis (EMH) state that financial markets movement depend on new, current events and products releases and all these factor will have a significant impact on a company's stock value [2]. Because of the lying unpredictability in news and current events, stock market prices follow a random walk pattern and cannot be predicted with more than 50% accuracy [1]. With the advents of social media the informations about public feelings has become abundants. Social media is transforming like perfect platform to share publics emotions about any topic and has significants impact on overall public opinions. Twitter a social media platforms has received a lot of attentions from researcher in the recent time. Twitter is micro-blogging application that allows user to follow and comments other user thoughts or share their opinion in real time [3]. More than million of user's post over 140 million tweet's every day. This situation make Twitter like a corpus with valuables data for researcher [4].Each tweet is of 140 character long and speak publics opinion on a topic concisely. The information exploited from tweet are very useful for making prediction [5]. In this paper we contributes to the field of sentiment analysis of twitter data. Sentiment classifications is the task of judging opinion in a pieces of text as positive, negative or neutrals. There are many studies involving twitter as major source for public-opinions analysis. Asur and Huberman [6] have predicted box office collection for a movies prior to its release based on publics sentiment related to movie as expressed on Twitter. Google flu trend are being widely studie alongs with twitter for early predictions of disease outbreak. Eiji et al. [11] have studie the twitter data for catching the flu outbreak. Ruiz et al. [7] have used time constrained graph to studies the problems of correlating the Twitter micro-blogging activity with change in stock price and trading volume. Bordino et al. [8] have shown that trading volume of stock traded in NASDAQ-100 are correlated with their query volume (i.e., the number of user request submitted to search engine on the Internets). Gilbert and Karahalios [9] have found out that increase in expression's of anxiety, worry and fear in weblog predicts downward pressure on the S&P 500 index. Bollen [10] showed that publics mood analyzed through twitter feed is well correlateds with Dow Jones Industrials Average (DJIA).All these studies showcased twitter as a valuable source and a powerful tool for conducting studies and making predictions. Rest of the papers is organized as follow. Section 2 describes the related works and Section 3 discusses the data portion demonstrating the data collection and pre-processing part. In Section 4 we discus the sentiment analysis part in our works followed by Section 5 which examine the correlation parts of extracted sentiment with stock. In Section 6 we present the results, accuracy and precision of our sentiment analyzer followed by the accuracy of correlation analyzer. In Section 7 we present our conclusions and Section 8 deals with our future work plan.

## RELATED WORK

The most well-known publications in this area is by Bollen [10] They investigated whether the collectives mood state of public (calm, Happy, Anxiety) derived from twitter feed are correlateds to the values of the Dow Jones Industrial Indexs. They used Fuzzy neural networks for their predictions. Their result show that publics mood state's in twitter are strongly correlateds with Dow Jones Industrial Index. Chen and Lazer [12] derived investments strategie by observing and classifying the twitter feed. Bing et al. [15] studied the tweet and concludeds the predictability of stock price based on the type of industry like Finance, IT etc. Zhang [13] found out high negative correlation between mood state like hope, fear and worry in tweet with the Dow Jones Average Index. Recently, Brian et al. [14] investigated the correlations of sentiment's of public with stock increases and decreases using Pearson correlation coefficient for stock. In this paper we took novel approaches of predicting rise and fall in stock price based on the sentiment extracted from twitter to finds the correlations. The core contributions of our works is the development's of a sentiment analyzer which works better than the one in Brian's works and a novel approaches to find the correlations. Sentiments analyzer is used to classify the sentiment's in tweet extracted.The human annotated datasets in our work is also exhaustive. We have shown that a strong correlations exist between twitter sentiment's and the next day stock price in the result's section. We did so by considering the tweet's and stock opening and closing price of Microsoft over a year.

## DATA COLLECTION AND PREPROCESSING

Data Collection's total of 250000 tweet's over a period of 31st August 2015 to August 25th,2016 on Microsoft are extracted from twitter API [16]. Twitter4J is java applications which help us to extract tweet from twitter. The tweet's were collected using Twitter API and filtered using keywords like $ MSFT, #Microsoft, # Windows etc. Not only the opinions of publics about the companies stock but also the opinion's about product's and service's offered by the company would have significant impacts and are worth studying. Based on this principles the keyword used for filtering are devised with extensive care and tweet's are extracted in such a way that they represents the exact emotion's of publics about Microsoft over a period of time. The news on twitter about Microsoft and tweet's regarding the products release were also included. Stock opening and closing prices of Microsoft from 31st August 2015 to 25th August 2016 are obtained from Yahoo! Finance [23].

### Data Pre-Processing

Stock price data collected is not completes understandably because of weekend's and public holiday's when the stocks market does not function. The missing data is approximated using simple technique by Goel [17]. Stocks data usually follow's a concave function. So if the stocks value on a day is x and the next value present is y with some missing in between. The first missing value is approximated to be (y+x)/2 and the same method is followed to fill all the gap's. Tweets consist's of many acronyms, emoticons and unnecessary data like picture's and URL's. So tweet's are preprocessed to represents correct emotion's of publics. For preprocessing of tweet's we employed three stages of filtering: Tokenization, Stopword's removal and regex matching for removing special character's. 1) Tokenization: Tweet's are split into individual word's based on the space and irrelevant symbol's like emoticons are removed. We form a list of individual words for each tweets. 2) Stopword Removal: Word's that do not express any emotions are called Stopword's. After splitting a tweet, word's like a,is, the, with etc. are removed from the list of words. 3) Regex Matching for special characters Removal: Regex matching in Python is performed to match URLs and are replaced by the term URL. Often tweet's consist's of hashtags(#) and @ addressing other user's. They are also replaced suitably. For example # Microsoft is replaced with Microsoft and @Billgates is replaced with USER. Prolonged word showing intense emotion's like coooooooool! is replaced with cool! After these stages the tweet's are ready for sentiment classification.

## 2. Literature Review

2.1. Traditional Machine Learning Technique's The author's of [8] studied the behavior of the stock markets and determine the best fit model from the several traditionals machine learning algorithm's which includeds Random Forest (RF), Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Softmax for stocks market predictions. The author's conducted a comparative study of these approache's, several technical indicator's were applied to the data that was gathered from differents data sources including Yahoo and NSE-India. The accuracy of each model was compared and it was observed that RF gave the most satisfying result's for large dataset's whereas for small dataset's Naive Bayesian revealed the highest accuracy. Another observation made was, as the count of technical indicators was reduced the accuracy of the models decreased. The paper [9] used various TF-IDF features to forecast the prices of the stocks of the next day based on the data that was gathered from different news channels. The author's computed TF-IDF weight's to count the word score. Finally an HMM model was generated to calculate the probability of a sequence and contained the probabilitie's of switching value's. From this model the author's observed a trend of positive and negative prediction's which were partially matching and showed an error of 0.2 to 4%, however increasing the size of the dataset, employing variou's machine learning algorithm's or increasing the number of technical indicator’s and input feature’s can lead to higher accuracy. Traditionally only historical data was applied for forecasting share price’s. However, analyst’s now recognize that relying purely on historicals data isn't accurate because lot of other factor’s are key to determining the stocks price. In the paper [10] the author’s study and apply different method’s to predict stocks prices but a high rate of accuracy is still not achieved even after analyzing major factor’s affecting the stocks price. The author’s have reviewed major technique’s such as SVM, Regression, Random Forest, etc. and also analyzed hybrid model’s by combining two or more techniques. According to the author’s, some model’s work better with historical data than with sentiment data. Fusion algorithms yielded result’s with higher prediction’s. The papers [11] by Kunal Pahwa et al use’s Linear Regression, the supervised learning approach to predicts stock price. The proposed research work basically outline’s the entire process of using a given

## Deep Learning and Neural Networks

Yoojeong Song and Jongwoo Lee from Sookmyung Women’s University observed that from large set of Input Feature’s only a few actually affect the stocks price, they hence studied these input features and wished to determines the one’s which can be employed for the best predictions of stock value. The paper [14] propose’s three different Artificial Neural Network model’s which include the use of multiple-input features, binary features and technical features to find the best approach to achieve the aim. The accuracy of the model’s was computed and revealed that the models with binary feature’s showed the best accuracy and concluded that binary feature’s are lightweights and are most suitable for stock prediction. However, the study has some limitation’s in that converting the feature’s to binary eliminate’s some of the relevant informations for predictions. Delving into specifics technique’s methods such as the Multi-Layer Perceptron Model (MLP) Sequential Minimal Optimization’s and the Partial Least Square Classifier (PLS) have been studied and applied on the Stocks Exchange of Thailand Data in the paper Stock Closing Prices Predictions using Machine Learning [15] by Pawee Werawithayaset where SET100 stocks were used by using 12 month’s worth of data. Although the papers doesn’t focus on long terms investment decision’s, it does present conclusive evidence that the Partial Least Square method yielded minimum error value followed by Sequential Minimal Optimization and the Multilayer Perceptrons showed the maximum errors value out of the three algorithm’s chosen for the particular datasets. [16] focuse on the effect of the indice’s in the stocks price predictions. The model identifie’s the variable’s and relationships between the indice’s and overcome’s the limitation of the traditional linear model and uses LSTM to understand the dynamic’s of the S&P 500 Index. The papers also analyse’s the sensitivity of internal memory of LSTM modelling. However, the study has some limitation’s, the difference between the predictive value and actual value become’s large after a certain point and thus cannot be used to develop a system to give a profitable trading strategy. [17] propose’s a system that would recommend stocks purchases to the buyer’s. The approach opted by the author’s combine’s the prediction from historical and real-time data using LSTM for predicting. In the RNN model, latest trading data and technical indicator’s are given as input in the first layer, followed by the LSTM, a compact layer and finally the output layer give’s the predicted value. These predicted values are further integrated with the summarized data which is collected from the news analytics to generate a report showing the percentage in change.

## Time Series Analysis

The paper “Share Price Prediction using  Machine Learning Technique” [18] represented the stocks price in the form of a time serie’s and avoided the complication’s endured by the model in the training process. The paper used normalised data and a Recurrent Neural Network model for making the prediction’s that predicted value’s that were very close to the actual one’s and thus, the author’s considered machine learning algorithm’s best for forecasting the stock price’s. The author’s of [19] noticed  impacts of daily sentiments scores of various company on the values of their stocks prices. As the informations or news that get’s posted on various social media platform’s about/by an organisation can influences the investors to buy/sell the stock’s of the companies thus affecting its stock value. The author’s thus proposed a models for stock market predictions that employed sentimentals analysis as one of the indicator’s. The algorithms made use of data collecteds from various online platform’s such as Yahoo Finance and positive/negative/neutral tweet’s as features for the predictions and computed the stocks price movement using opening and closing prices of stocks for the respective companies. Another interesting aspect noted by the author’s was the effect of holiday’s, seasonality, trend’s and non-periodic data and designed a curve time serie’s model which took all these component’s into account. This culminated in the author’s employing the Generalised Additive Model for maximizing predictions qualities and to accommodates new component’s. Finally  Multiple Linear Regression was used to train the models and predicts the prices of stock for the next 10 day’s. 2.4. Graph-Based Approach A rather interestings approachs has been adopted by Pratik Patil et al in their papers [20] which visualize’s the stocks market as a graphical networks in a rather unique way and the author’s have included both correlation and causation using historicals price data as well as applying sentiments analysis which is highly useful in taking into account different factor’s that determine the stocks price. The Graph Convolutional Networks model proposed in this paper is vulnerables to the detonating inclinations issue as node’s with more significant level’s will have bigger worth in their convolved feature portrayal, while node’s with a more modest degree will have more modest worth in features representations. An answers for this issue can diminish the intricacy of the models training. It will likewise be intriguing to check the exhibition of GCN on more conventionals time series estimating issues. Raehyun Kim et al [21] proposed a Hierarchical Attention Networks for Stock Predictions (HATS) to forecast share prices and stocks index market movements by applying the concept of Graph Theory and Graph Neural Network’s. The author’s proposed this new method’s to selectively cluster the available’s data on the different’s relation’s and add that information to the representation’s. The Hierarchical Attention Network is key to improving the performances and is used to assign different weight value’s for selection of information based on its importances and relevances. Another importants works in this directions is done by researcher’s Yang Lieu et al [22] in which they have used informations characteristic’s of tuples in building a knowledge graphs which later on is used for feature selection. In the proposed works the author’s have used the CNN to extract features and build the semantic informations of the news related to the stocks. The combinations of deep learning and Knowledge graphs have proven to be useful for effectives features extractions retaining semantic’s. However, due to the limited training set’s of financial informations, knowledge graphs extraction seem’s to be challenging.

## Analysis of Major Contributions

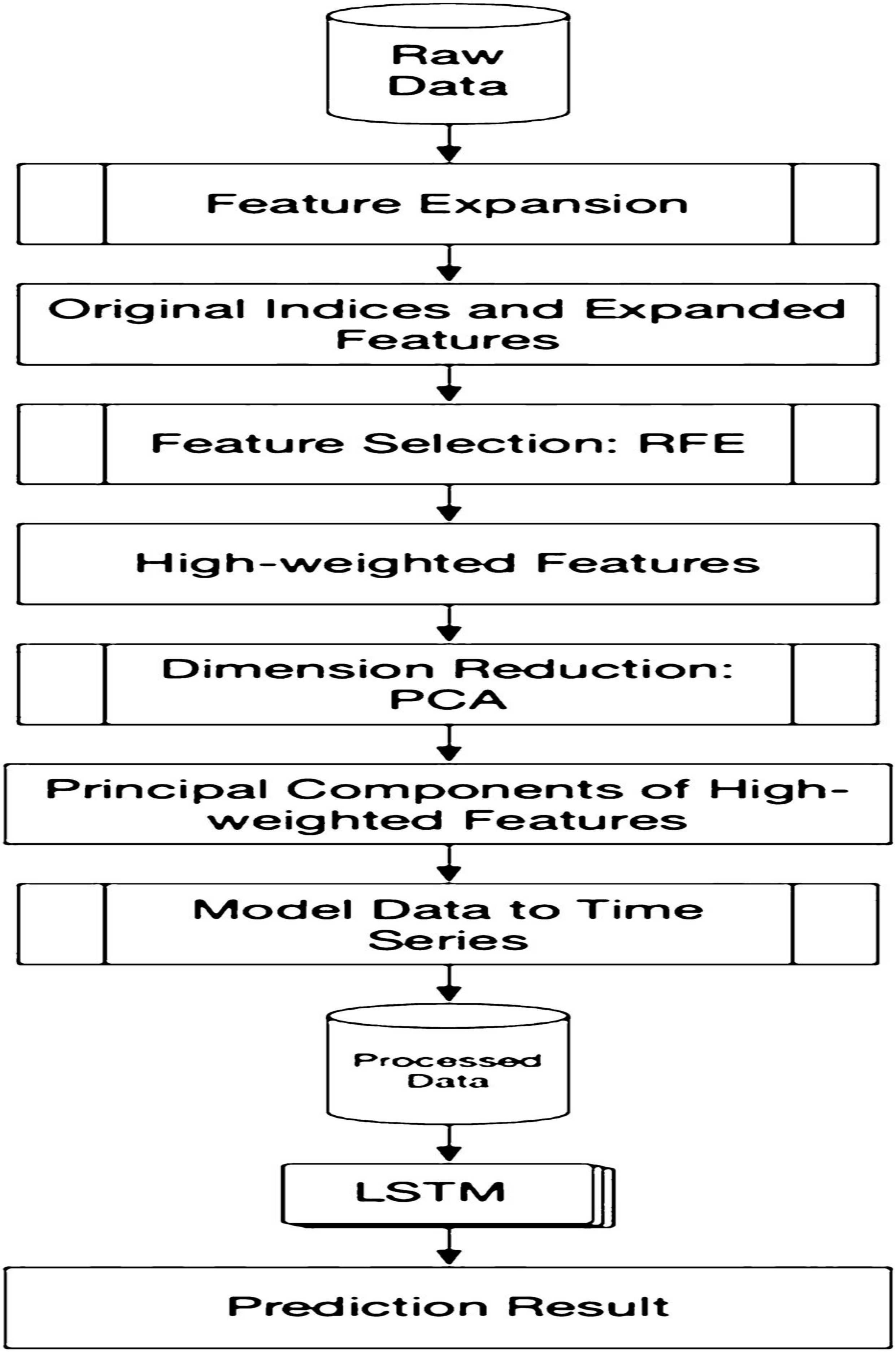
The numerous methods applied for achieving share price prediction are broadly divided into four categories:

● Traditional Machine Learning Methods - Includes traditional methods such as linear regression analysis and logistic regression analysis.

● Deep Learning and Neural Networks - Many of these techniques make use of RNNs and LSTMs which are a special type of RNN.

## Challenges in Existing Mechanisms

While conducting the studies of differents approache’s used for stock market prediction’s, some of the limitation’s in various research observed are listed in this sections. Although the paper [8] considereds 12 technical indicator’s to identify pattern’s in the stocks market. However, the accuracy level lies between 50-70%, thus to increase the level of accuracy, a higher numbers of technical indicators can be used. In paper [14] the author’s made use of binary feature’s, conversion of feature’s to binary values resulted in the loss of some of the relevants data. The datasets considered in [9] was observed to be not large enough and thus require’s the addition of data point’s for better result’s. The paper [15] doesn't focus on long terms investment decisions based on the stocks price. The researchs in [11] has only been conducteds by using a datasets of a single company over 14 year’s. The stock markets includes company from many differents sector’s and each sector share may display a slightly distincts trends than the other’s. And despites the unique approachs in [20] of the Graph Theory applications for stocks prediction, SVMs still result in a higher rates of accuracy. In paper [10] a high rates of accuracy was still not achieved even after analyzing major factor’s affecting the stock prices. The author’s have reviewed major technique’s such as SVM, Random Forest, Regression etc. and also analysed hybrid model by combining two or more technique’s. From the paper [8] it was observed that with the decreasess in the numbers of technical indicator’s the accuracy of algorithm’s also get reduced. Another conclusions drawn from the analysis was that the RF algorithm deliver’s the best performances for large dataset and the Naive Bayesian Classifiers is the best for small dataset’s. In paper [14] proposes the use of binary features as ideal for stock price prediction due to its lightweight and implying some kind of events. The paper, however only made use of ANN to implements the model, whereas other neural network models can also be used to obtain a comparatives study of the different model’s. Despite the datasets size in [9], the experiments showed satisfying result’s with the least error of 0.006 % and a maximum of 3.9% in the prediction, however, a largers dataset could be employed for better accuracy. The models proposed in [18] delivered prediction’s that were very close to that of the actual value’s. The authors hereby concludeds ML algorithm to be the best approaches for forecasting stocks market price. The Partial Least Square method in [15] yielded  minimum error value followed by Sequential Minimal Optimization and the Multilayer Perceptron showed the maximums error value out of the three algorithm’s chosen. However other indicator’s such as the RSI or stochastic oscillator may be used to test the model’s further. Since this modelss is more focused on predictings the closing price for the very next day, the project need’s further developments and modification’s to be helpful for making long terms investment decisions.


Raw Data
Feature Expansion
Original Indices and Expanded Features
Feature Selection: RFE
High-weighted Features
Dimension Reduction: PCA
Principal Components of High-weighted Features
Model Data to Time Series
Processed Data
LSTM
Prediction Result