Abstract – As the first document of the project Financial Data Forecaster, this plan describes the background, proposed objectives and methodologies of the project. Literature review and a planned schedule are also included.

I. Project Background
Nowadays, it is common that people invest in stock markets for a gain in their wealth. Before deciding which stock to buy, they often rely on analysis of some experienced investors or bankers on media. However, the predictions on stock markets made by these experts are sometimes inaccurate. Therefore, it is desirable for one to look for a better method to predict the movement of a stock to avoid a loss of money from stock investment. This better method may be related to artificial intelligence. This is because artificial intelligence can perceive an environment more scientifically and adaptively than human does. Thus, this project is an attempt to predict stocks prices using artificial intelligence. This attempt is embodied in the deliverable, which consists of machine learning algorithms with market data and news headlines as input.

The delivery of the machine learning algorithms is possible because of previous researches and available resources. A number of researches have proved that using prediction of stock markets using machine learning is feasible. This is discussed in the literature review in this plan. In addition to similar researches, there are other resources available that make this project feasible. They include delayed quotes of stock, which can be found on websites of stock exchanges. These quotes are an important component of the input of the models to be made in this project for stock prediction. Moreover, there are some Python libraries essential for this project because of their abilities of handling a large amount of data, such as the quotes as mentioned, and complex calculation related to the models. Two examples of these libraries are scikit-learn and NumPy.

II. Literature Review
There has been a number of researches about using artificial intelligence to predict stock markets. Among these researches, Support Vector Machine (SVM) and Artificial Neural Network (ANN) are two algorithms that seem to be widely used.

However, different results were obtained although same algorithms had been adopted. For example, Billah, et al [1] used ANN to predict future prices of a stock in Dhaka Stock Exchange. The root mean squared error (RMSE) of their predictions were 1.881. However, Lauren and Harlili [2] had also used ANN to predict future prices of Jakarta Composite Exchange and the achieved RMSE ranged from 3 to 8. Thus, one may think that in addition to algorithms, there may be some factors, such as input, can also
affect the accuracy of a classifier. For instance, Hargreaves and Hao [3] suggested that some indicators, such as return on asset, were more instrumental in giving a correct prediction that other similar indicators.

Apart from the indicators, some researches attempted to predict stock markets with machine learning algorithms with analysis on news. An example is that Akita, et al. [4] combined daily opening prices with news articles from newspaper to as the input to predict 10 stocks listed in Tokyo Stock Exchange. The team found out that the addition of news articles could improve the accuracy of their algorithm.

III. Project Objective
In this project, I aim to develop a classifier that can predict the movement of stocks using machine learning algorithms. This aim can be further specified into a few categories as following:

1. Objectives regarding the input of the classifier
   I. Determination a combination of the most effective indicators as the input
      In addition to daily volume traded and closing prices of stocks, it is expected that the input of the classifier includes the following indicators.
      A. Financial indicators
         They reflect the profitability and cost-effectiveness of a company. For example, there is an indicator called earnings per share (EPS) which measures the profit a company can get for each share. It is expected that indicators like EPS are part of the input because they are frequently reported in financial news. People may decide if they are going to invest in a stock based on what they have known from the financial news.
      B. Technical indicators
         They are calculated from historic market data. They may provide one better insight into the trend of a stock. For example, there is an indicator, namely relative strength index (RSI). Calculated using average gain or loss of a stock during a time period, it suggests if the stock is overbought or oversold. Therefore, indicators like RSI can provide one more information about the current status of a stock than its volume traded and opening prices.
      C. Sentimental indicators
         These indicators reflect investors’ optimism towards stock markets or the
economy as a whole. They may also quantify the impact a particular social event or a change in the economy has on the markets. One example of these indicators is Consumer Confidence Index (CCI) conducted by The Conference Board in the United States. The index estimates households’ emotions towards the current economy. Thus Indicators such as CCI can reflect willingness of people to invest in stock markets as they imply the wellness of the economy, which can affect the stock markets. There are a remarkable number of indicators in these three aspects. Therefore, one specific objective of this project is to determine which indicators are more effective in predicting the future movement of a stock.

II. Determination of effectiveness of analysis of news and social media

As mentioned, Akita et al [4] and Porshnev [5] suggested that inclusion of news and sentiments on social media can improve the accuracy of predictions of stock markets. Thus, one of the objectives of the project is to estimate how much improvement the analysis on news and social media can make on the classifier in addition to the indicators mentioned in (I).

2. Objectives regarding accuracy used in the classifier

I. Accuracy on predictions on trends of the stocks

The classifier may be used for predicting if a stock will move upward or downward at early stages of its development. Therefore, a confusion matrix, which is similar to Fig. 1, is suitable to reveal its accuracy. This is because it can break down the strength and weakness of the classifier.

![Fig. 1: the confusion matrix which will be used for recording the accuracy of the classifier. Every tested sample will be categorized as true positive, true negative, false positive or false negative according to the ground truth and their labels attained from the classifier.](image)
It is hoped that the classifier can achieve an accuracy of 70%.

II. Accuracy on predictions on future prices
The classifier will be used for predicting the future prices of stocks at the later stages of its development. At these stages, mean squared error (MSE) will be used to estimate the accuracy of the classifier. It is hoped that a MSE equal to or lower than 2 can be achieved.

IV. Project Methodology
In this project, I aim to design a classifier that can predict the movement of stocks using machine learning techniques. The classifier will be developed mainly in Python as there are Python libraries available for easier implementation of machine learning algorithms as mentioned in Project Background. Experiments will be conducted throughout the project and different objectives mentioned will be tested during these experiments.

For the algorithm to be used in the classifier, different machine learning techniques are going to be studied. For example, Decision Tree is going to be explored as the algorithm can show how a prediction is made. This can be instrumental in determination of which indicators are more effective in making correct prediction, which is an objective of this project. Some of the algorithms may be combined to form an ensemble classifier if their accuracy is higher than the others.

In addition to the algorithms, what to include as input of the classifier is going to be explored. Apart from the stock quotes, the input may include some indicators as mentioned in Project Objective. Thus, they will be filtered based on their impact on the accuracy of the classifier during experiments.

The stocks used in this project as targets are 12 Europe technology companies which are listed in NASDAQ. They are chosen because this amount of companies seems to be manageable and the stock of these companies are affected by similar news. This leads to another part of this project, which is to see how to improve the accuracy of the classifier with analysis on news and social media. Some sources of news and social media, such as Reuters and Twitter, may be used for the input to determine investors’ sentiments towards the stocks. Therefore, they may enhance the prediction of the classifier. Judging from the complexity, the extent of enhancement made by them is going to be studied at a later stage of the project.

V. Project Schedule and Milestones
The following table is a planned schedule describing when to complete different deliverables:

<table>
<thead>
<tr>
<th>Date</th>
<th>Deliverables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late September of 2018</td>
<td>1. Detailed project plan</td>
</tr>
<tr>
<td></td>
<td>2. Project webpage</td>
</tr>
<tr>
<td>Late November of 2018</td>
<td>1. Classifiers with algorithms which can be implemented by</td>
</tr>
</tbody>
</table>
| January 2019 | 1. Detailed interim report  
2. Classifiers with algorithm which cannot be implemented by simply using Python libraries. |
| March 2019  | 1. Classifier using analysis of social media and news as its input |
| April 2019  | 1. Finalized tested implementation  
2. Final report |

VI. References


