Stock Price Prediction using Linear Regression based on Sentiment Analysis

Yahya Eru Cakra
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
yahya.eru@ui.ac.id

Bayu Distiawan Trisedya
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
b.distiawan@cs.ui.ac.id

Abstract*— Stock price prediction is a difficult task, since it very depending on the demand of the stock, and there is no certain variable that can precisely predict the demand of one stock each day. However, Efficient Market Hypothesis (EMH) said that stock price also depends on new information significantly. One of many information sources is people's opinion in social media. People's opinion about products from certain companies may determine the company's reputation and thus affecting people's decision to buy the stock of the company. When using opinion as primary data, it is necessary to make a suitable analysis of it. A famous example using opinion as data is sentiment analysis. Sentiment analysis is a process to determine emotion/feeling within people opinion about something, in this case products of some companies. There are some researches about sentiment analysis used to predict the stock prices. Bollen on his research concludes that people opinion on social media such as Twitter can predict DJIA value with 87.6% accuracy. This shows that there is a relation between sentiment analysis and stock prices. Our purpose on this research is to predict the Indonesian stock market using simple sentiment analysis. Naïve Bayes and Random Forest algorithm are used to classify tweet to calculate sentiment regarding a company. The results of sentiment analysis are used to predict the company stock price. We use linear regression method to build the prediction model. Our experiment shows that prediction models using previous stock price and hybrid feature as predictor gives the best prediction with 0.9989 and 0.9983 coefficient of determination.

Keywords*— *linear regression, sentiment analysis, stock price, supervised learning, Twitter

I. INTRODUCTION

Stock price prediction is a difficult task. It is because there is no certain variable that can precisely predict the stock price every day. Based on Efficient Market Hypothesis (EMH), new information is a significant factor that effects changes of stock price [1]. This information, such as news about company can influence people decision whether or not they will buy the company's stock. More people buy the company's stock, the price are getting higher. People tend to buy a company with good reputation. One way to know company's reputation is by seeing relationship between the company and customer [2]. The explosion of social media usage force many companies to create their official account in social media in order to keep in touch with their customer. This make customer can express their opinion about products easily. One of the social media that commonly used by company is Twitter.

There are several researches about how the information from social media can affects the stock price. Based on research conducted by Johan Bollen, et.al[3], it concluded that certain mood states of Twitter data can predict the Dow Jones Industrial Average (DJIA) value with 87.6% accuracy. Another research conducted by Anshul Mittal and Arpit Goel [4], shows that with the DJIA value, calmness and happiness mood states of twitter data on previous days can predict the DJIA value on the current day with 75.56% accuracy. This shows that information from Twitter can really be used to predict stock data.

Indonesia is the 5th country with highest number of Twitter active user, especially Jakarta where 2.4% of all Twitter post comes from [5]. Since the previous research mentioned was using English, author was curious about the effects of Twitter data to stock price of Indonesian Company. This problem gives us motivation to conduct this research.

The contribution of this research lies in the use of existing classification and prediction algorithm to the dataset. The dataset consists of twitter dataset and stock price dataset. Twitter dataset used was in Bahasa and stock price dataset retreived from several companies in Indonesia.

II. DATASET

This research used two kinds of data. Stock prices of several companies in Indonesia and data contained opinions about certain products produced by the companies mentioned. Opinions was shared via Twitter. Companies that were being chosen were companies with fluctuating stock prices and it's products already popular in Indonesia. Companies being research were as follows:

TABLE I. COMPANY LIST

Companies List
XL Axiata
Fast Food Indonesia (KFC)
Indosat
Unilever Indonesia
Tempo Intimedia
Mayora Indah
Pembangunan Jaya Ancol
Indofood Sukses Makmur
First Media
Bank Mandiri (Persero)

978-1-5090-0363-1/15/$31.00

Bank Central Asia.
MNC Investama
Bank CIMB Niaga

Data gathering had been done in two weeks, from April 14th 2015 to April 30th 2015.

A. Tweet Dataset

Tweet dataset was gathered using Twitter REST API. Type of the API used was Search API. Keywords used for query parameter were company's official Twitter account and names of the products produced by the companies. Tweets retrieved were in Bahasa. Retrieved data was JSON formatted. For calculating the sentiment analysis purpose, we only collect tweet id, posting time, user who posted the tweet and tweet/status.

B. Stock Price Dataset

Stock price dataset was gathered using Yahoo Finance CSV API. Information being collected were the open stock price and close stock price of the companies for each day. Retrieved data was CSV formatted. This data were then being used as the predicted value, also combined with the result of sentiment analysis to create prediction model.

III. Sentiment Analysis

Sentiment analysis is a process to classify the polarity of opinion. One important thing to determine polarity is sentiment lexicon. Sentiment lexicon is a word or phrase containing emotion/feel of a sentence. There are two kinds of sentiment lexicon. They are positive lexicon and negative lexicon. Positive lexicon is word or phrase which expresses positive emotion/feeling (e.g. good, awesome, delicious, etc) and negative lexicon is word or phrase which expresses negative emotion/feeling (e.g. bad, ugly, sick, etc). In this research, lexicon sentiment retrieved from gathered data. First the gathered data were being tokenize into single word. Then, using the definition from Indonesian dictionary, formal word that categorized as lexicon was chosen. As for the informal word, they were manually checked by looking at the word similar to it. Then the positive lexicon and negative lexicon were separated.

There are many kind of classification in determining polarity of opinion. There are classifying opinion into 6 mood states (Calm, Alert, Sure, Vital, Kind and Happy), classifying opinion into 3 classes (positive, negative, neutral), etc. One of many methods that can be used in sentiment analysis is classifying by supervised learning. The method classifies opinion based on information/features that can be extracted from the data [6]. Besides lexicon, there are other features that can be extracted and used in classification. Some of them are:

- Words and it's weight

Calculating the weight of words in document will make it easy to determine which word is important and which is not.

- (Part of Speech) POS Tagging

POS tagging is usually used to determine type of word in a sentence. By knowing the type of word, It is easy to choose the important information. Type of word that are usually being used are noun and adjective.

- Sentiment shifters

Sentiment shifters are a word or phrase that reverses the emotion/feel of a sentence (e.g. not, don't, etc).

There are many algorithms that can be used in supervised classification. In this research used algorithm are Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Random Forest and Neural Network (with single layer perceptron).

A. Support Vector Machine (SVM)

SVM is an instance based algorithm which created a linear function which maximizes the distance between classes [7]. The algorithm uses instance data on the edge of the class to create the class function. This instance data is called support vector. Below is the illustration of SVM.

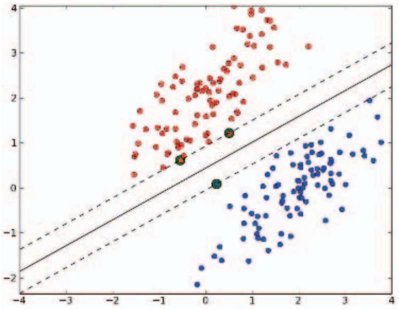

Fig. 1. Illustration of SVM (Source: www.mblondel.org)

The linear line is the classifier function. Black points represent instances that are used to create the function and are called support vectors. Distance between the dashed line is called margin. The aim of SVM algorithm is to construct a function that maximize margin.

B. Naïve Bayes

Naïve Bayes is a classification algorithm with statistical approaches. It assumes that every feature of instance data is independent [8]. The algorithm uses conditional probability to determine which class is belongs to each instance data. Below is the probabilistic function in Naïve Bayes.

$$P(C|F_1,F_2,F_3,\ldots F_n)=\frac{P(F_1|C)P(F_2|C)P(F_3|C)\ldots P(F_n|C)P(C)}{P(F_1,F_2,F_3,\ldots F_n)} \quad (1)$$

- $P(C|F_1,F_2,F_3, \ldots F_n)$ is probability of one instance classified as C given feature combination $F_1,F_2,F_3, \ldots F_n$
- $P(F_1|C)$ is probability of feature F_1 given data classified as C
- P(C) is probability of data classified as C
- $P(F_1,F_2,F_3, \ldots F_n)$ is probability of feature combination $F_1,F_2,F_3, \ldots F_n$

The instance data is being classified into class with higher probability value.

C. Decision Tree

Decision Tree is a tree based classification algorithm which represents a form of a tree. There are three part of a tree. Root node, internal node and leaf node [9]. Root and internal node represent the feature of instance data whereas leaf node represents the class. Each split of root or internal node represent values of the feature represented by the split node. Below is the illustration of decision tree.

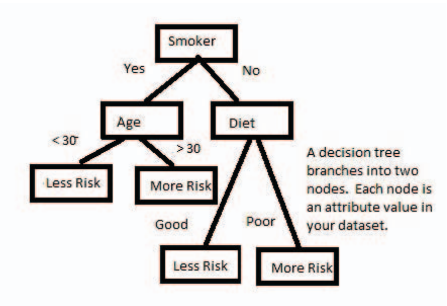
Smoker
Yes
No
Age
Diet
< 30
> 30
Less Risk
More Risk
Good
Poor
Less Risk
More Risk
A decision tree branches into two nodes. Each node is an attribute value in your dataset.

Fig. 2. Illustration of decision tree (Source: http://www.refactorthis.net/)

Problem that often comes up is determining which feature that are being used as the root and internal nod for each depth. One that can be used is by choosing feature with high information gain. Information gain is a value ranged from 0 to 1 which represent how significant a feature classifying the data. The more information gain value close to 1 is the more significant a feature is. Information gain can be calculated using below function.

$$InfoGain(C,f) = Entropy(C) - Entropy(C|f) \quad (2)$$

- C is class
- f is feature which information gain value is going to be checked

Below is function to calculate class entropy value.

$$Entropy(C) = -\sum_{j=1}^{n} prob(j)\log_2 prob(j) \quad (3)$$

- C is class
- n is number of classes
- j is values of class

Below is function to calculate entropy of one feature

$$Entropy(C|f) = \sum_{j=1}^{n} prob(f=j)Entropy(K|f=j) \quad (4)$$

- C is class
- n is number of value in the feature
- j is feature values

D. Random Forest

Random Forest is an ensemble classification method which consists of many tree based model classifier. It constructs each tree by splitting number of features for each split without pruning [10]. Thus, it needs two parameters. They are number of tree and number of feature in every splitting. The features are chosen randomly with replacement. After number of random tree are constructed, each data will classified by all of the trees. Tree which classifies most instance data correctly will be used as the main classifier. This is called voting.

E. Neural Network

Neural network is a classification algorithm which inspired by human neural network. This method is quite similar to classification algorithms which use linear approach. The algorithm creates a function which calculates weight for each feature [11]. Below is the linear equation for Neural Network.

$$x = f_0 w_0 + f_1 w_1 + f_2 w_2 + \cdots + f_n w_n \quad (5)$$

- f is feature
- w is weight of each feature

There are many methods in determining value of w. One of which is gradient descent. It is an iterative method which updates value of w by minimizing the value of square error. Below is the function of gradient descent.

$$W_t = W_0 - \eta \frac{\partial E^2}{\partial w} \quad (6)$$

- W_t is updated weight
- W_0 is initial weight
- η is learning rate
- E^2 is function of square error

If the learning rate value is high, the appropriate weight will be reached quickly but the function is more unstable. In the other hand, if the learning rate value is low, the appropriate weight will be reached slowly but the function is more stable. Since the output of the created function is continuous, another function is needed to transform the result into discrete. This function is called activation function. Examples of activation function are sign function, sigmoid function, etc.

IV. Prediction Model

There are three things that were predicted in this research. They are price fluctuation prediction, margin percentage prediction and price prediction. Price fluctuation prediction used classification by supervised learning method. Margin percentage and price prediction used linear regression, since the value that was being predicted was not categorized.

Linear regression is one of regression method to be used for classifying numerical class [12]. It creates linear function by calculating weight values (w) for each feature (β). The function can be seen as follow:

$$x = \beta_0 w_0 + \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_n w_n \quad (7)$$

X are regression value for one instance data. To have a clear understanding about linear regression, here is the illustration of linear regression.

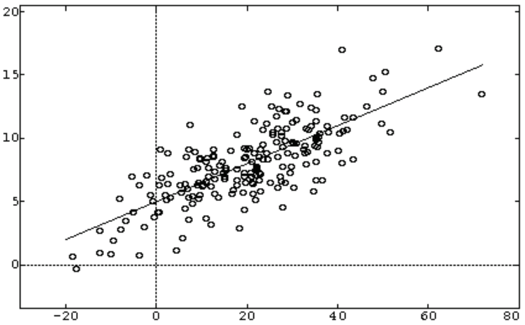

Fig. 3. Illustration of Linear Regression (Source: en.wikipedia.org)

As seen in picture above, there is a linear line that represents the distribution of data. Distance from each instance data to the linear line is called error or residual. The linear function is created by searching appropriate weight value for each feature to minimize error of each instance data (mean error).

There are many ways to evaluate a linear regression model. Model which fitted most data has a normal distribution in it's residual. Besides, it also can be evaluated by looking at the coefficient of determination value (R^2). R^2 is square of coefficient of correlation (R) between predicted values and actual values. R^2 ranged from 0 to 1. The more R^2 value close to 1 is the more data that are fitted by the model.

V. Experimental Design

A. Sentimen Analysis Design

The purpose is to create models that can classify the polarity of tweet data. Method used was classification by supervised learning. In this research, the gathered Tweets were being classified into three classes (positive, negative and neutral). Features that were used in classification were sentiment lexicon and sentiment shifters which retrieved from the tweets which already gathered. Each tweet will be examined, whether it contained positive lexicon, negative lexicon or none of both. It is also being checked about the existence of sentiment shifters.

Several algorithm used in classification were those which explained in chapter III. Two classification models with highest accuracy was later used to classify all the tweets into three classes. After all tweets were classified, the percentage of positive tweet was calculated for each day. The calculation formula were as follows:

$$\% = \frac{\#positive\_tweet}{\#positive\_tweet + \#negative\_tweet}$$

The neutral tweets was not used in this research, because most of which were spam tweets or promotion tweets which was not relevant to the research.

B. Prediction Model Design

1) Price Fluctuation Prediction

The purpose is to predict whether the stock price of one company at observation day would go up or down compared to the stock price on the previous day. This prediction model used dominant tweet percentage on previous day (ranged 1 to 5 days) as the predictor.

It used supervised classification method with classification algorithm such as Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF) and Neural Network (NN) (with single layer perceptron). Created model was then tested to predict the price fluctuation of stock price based on tweet data that already classified by two sentiment classification model.

2) Margin Percentage Prediction

The purpose is to predict the stock margin percentage of one company at observation day using information from previous days as the features. The prediction model was created using linear model. The algorithm being used was linear regression (multinomial). There were three linear model created for this prediction. Each used different feature. Below is the linear model representation of the three models.

1. $M_x = w_0 + \sum w_i M_{x-i}$
2. $M_x = w_0 + \sum w_i T_{x-i}$
3. $M_x = w_0 + \sum w_i M_{x-i} T_{x-i}$

- M_x : margin percentage on day x
- T_x : positive tweet percentage on day x
- w_0 : intercept weight
- w_i : feature weight

The models used features from 1 to 5 days before observation day. For model with positive tweet percentage, it being tested with two kind of tweets data classified with two ways as explained before. The models then being evaluated based on the value of coefficient of determination (R^2)

3) Stock Price Prediction

The purpose is to predict the stock price of one company at observation day using information from previous days as the features. The prediction model was also created using linear model. The algorithm being used was linear regression (multinomial). There were three linear model created for this prediction. Each used different feature. Below is the linear model representation of the three models.

1. $P_x = w_0 + \sum w_i P_{x-i}$
2. $P_x = w_0 + \sum w_i P_{x-i} T_{x-i}$
3. $P_x = w_0 + \sum w_i H_{x-i}$
   $H_x = P_x + \text{sign}(T_x) D_x$
   $\text{sign}(T_x) =$

- P_x : stock price on day x
- T_x : positive tweet percentage on day x
- H_x : hybrid feature on day x
- D_x : $P_x - P_{x-1}$
- w_0 : intercept weight
- w_i : feature weight

The models used features from 1 to 5 days before observation day. For model with positive tweet percentage, it being tested with two kind of tweets data classified with two ways as explained before. The models then being evaluated based on the value of coefficient of determination (R^2).

VI. Result

A. Result of Sentiment Analysis

Sentiment analysis classification used five algorithms which are mentioned before to create the classification model. The accuracy of the classification model for each algorithm can be seen in table below.

TABLE II. Accuracy of Algorithm for Classification

Algorithm	Accuracy
Support Vector Machine (SVM)	38.64%
Naïve Bayes	**56.50%**
Decision Tree	46.02%
Random Forest	**60.39%**
Neural Network	42.52%

The result shows that highest accuracy was produced by classification model using Random Forest with 60.39% accuracy. The second highest accuracy was produced by classification model using Naïve Bayes with 56.50% accuracy. The two models then used to classify all the Tweets into three classes.

B. Result of Prediction Model

1) Price Fluctuation Prediction Model I

The accuracy of this model for each algorithm can be seen in below table and diagram.

TABLE III. Accuracy of Price Fluctuation Prediction (Data Classified by Naïve Bayes) (%)

Classifier	Number of Previous Days				
	1	*2*	*3*	*4*	*5*
SVM	66.34	66.34	66.34	66.34	66.34
NB	66.34	66.34	66.34	61.39	57.43
DT	66.34	66.34	66.34	66.34	66.34
RF	66.34	66.34	66.34	**67.37**	65.35
NN	66.34	66.34	66.34	59.41	65.35

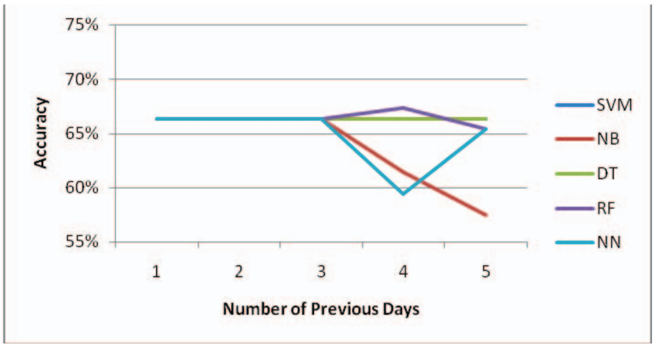

Fig. 4. Accuracy of Price Fluctuation Prediction Diagram (Data Classified by Naïve Bayes)

The result shows that highest accuracy was produced by prediction model using Random Forest algorithm and data from four previous days with 67.37% accuracy.

2) Price Fluctuation Prediction Model II

The accuracy of this model for each algorithm can be seen in below table and diagram.

TABLE IV. Accuracy of Price Fluctuation Prediction (Data Classified by Random Forest) (%).

Classifier	Number of Previous Days				
	1	*2*	*3*	*4*	*5*
SVM	66.34	66.34	66.34	66.34	66.34
NB	66.34	66.34	56.44	56.44	56.44
DT	66.34	66.34	66.34	66.34	66.34
RF	66.34	66.34	60.40	60.40	60.40
NN	66.34	66.34	61.39	61.39	61.39

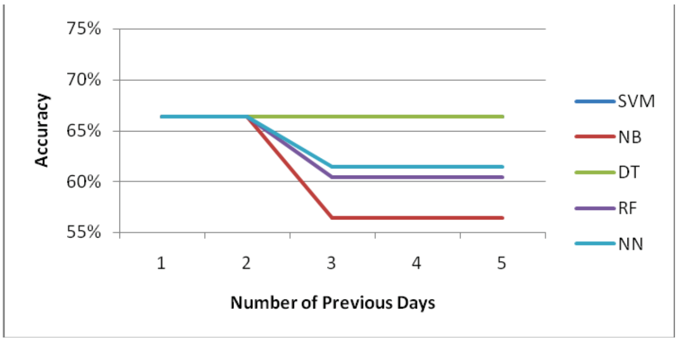

Fig. 5. Accuracy of Price Fluctuation Prediction Diagram (Data Classified by Random Forest)

The result shows that highest accuracy was produced by prediction model using all algorithm and data from one previous days with 66.34% accuracy. It also shows that the accuracy of the model tend to decrease with increasing of number of previous day.

3) Margin Percentage Prediction Model

The R^2 value of this model for each feature used can be seen in below table and diagram.

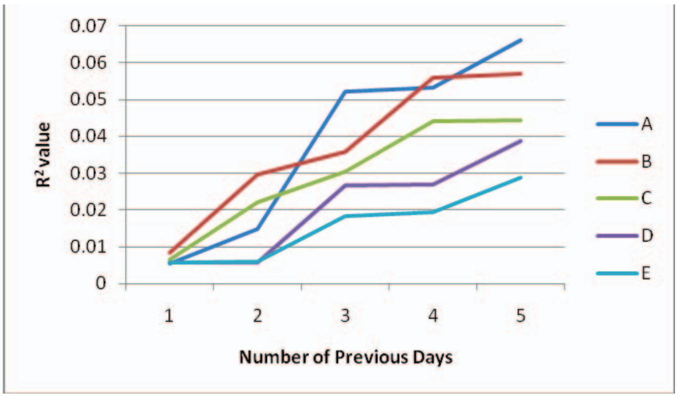

Fig. 6. Graphic of R^2 value of Margin Percentage Prediction Model

TABLE V. R^2 VALUE OF MARGIN PERCENTAGE PREDICTION MODEL

Features	Number of Previous Days				
	1	*2*	*3*	*4*	*5*
Margin percentage (A)	0.0054	0.0149	0.0521	0.0534	**0.0662**
Positive tweets percentage (tweet classified by Naïve Bayes) (B)	0.0085	0.0298	0.0359	0.0558	0.0570
Positive tweets percentage (tweet classified by Random Forest) (C)	0.0065	0.0221	0.0303	0.0441	0.0444
Margin percentage . Positive tweets percentage (tweet classified by Naïve Bayes) (D)	0.0058	0.0058	0.0268	0.0269	0.0388
Margin percentage . Positive tweets percentage (tweet classified by Random Forest) (E)	0.0058	0.0060	0.0185	0.0196	0.0289

The result shows that highest R^2 value was produced by prediction model using margin percentage feature on five previous days, that is 0.0662. However, all created model produced R^2 values which close to 0.

4) Price Prediction Model

The R^2 value of this model for each feature used can be seen in below table and diagram.

TABLE VI. R^2 VALUE OF STOCK PRICE PREDICTION MODEL

Features	Number of Previous Days				
	1	*2*	*3*	*4*	*5*
Price (A)	0.9984	0.9986	0.9986	0.9987	**0.9989**
Price . Positive tweets percentage (tweet classified by Naïve Bayes) (B)	0.8360	0.9078	0.9202	0.9217	0.9262
Price . Positive tweets percentage (tweet classified by Random Forest) (C)	0.9047	0.9288	0.9347	0.9348	0.9385
Hybrid feature (positive tweets percentage classified by Naïve Bayes) (D)	0.9978	0.9979	0.9979	0.9982	**0.9983**
Hybrid feature (positive tweets percentage classified by Random Forest) (E)	0.9978	0.9979	0.9979	0.9982	0.9983

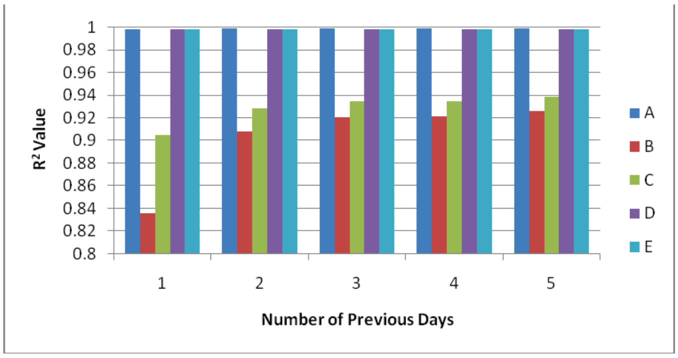

Fig. 7. Graphic of R^2 value of Stock Price Prediction Model

The result shows that highest R^2 value was produced by prediction model using price feature on five previous days, that is 0.9989. It is followed by prediction model using hybrid feature on five previous days that is 0.9983. All created models produced R^2 values which close to 1.

VII. CONCLUSION AND FUTURE WORK

From the result of this research, there are several things that can be concluded.

- Created sentiment analysis model using Random Forest algorithm can classify tweet data with 60.39% accuracy, and the one with Naïve Bayes algorithm can classify tweet data with 56.50% accuracy.
- In price fluctuation prediction, created models can predict whether the upcoming price will go up or down with highest accuracy of 67.37% for tweets data classified by Naïve Bayes and 66.34% for tweets data classified by Random Forest.
- In margin percentage prediction, created models have R^2 value which close to 0. It means that created models fitted only few data.
- In price prediction, created models have R^2 value which close to 1. It means that created models fitted lots of data. The highest R^2 value retrieved from model with previous price as the feature.
- Percentage of positive tweets decreases the R^2 value of a model.

There are also several suggestion that might be applied for future work.

- Improving sentiment analysis model by applying other feature in classification. Such as POS tagging, word weighting, etc.
- Expanding the period of data gathering into a month or more.
- Use another method to create prediction model which not linear.

REFERENCES

[1] Investopedia. (n.d.). *Efficient Market Hypothesis: Is The Stock Market Efficient?* Retrieved June 24, 2015, from Investopedia: http://www.investopedia.com/articles/basics/04/022004.asp

[2] Bracey, L. (n.d.). *The Importance of Business Reputation*. Retrieved Juli 9, 2015, from Business in Focus: http://www.businessinfocusmagazine.com/2012/10/the-importance-of-business-reputation/

[3] Bollen, J., Mao, H., & Zeng, X. J. (2010). Twitter mood predicts the stock market. *arXiv* .

[4] Mittal, A., & Goel, A. (2009). Stock Prediction Using Twitter Sentiment Analysis. *CiteSeerX* .

[5] Berita 8. (2013, November 21). *Ini 5 Negara Pengguna Aktif Twitter Terbanyak di Dunia*. Retrieved June 25, 2015, from Berita 8: http://www.berita8.com/berita/2013/11/ini-5-negara-pengguna-aktif-twitter-terbanyak-di-dunia

[6] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Claypool Publishers.

[7] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed).* Burlington: Morgan Kaufmann, pp. 191-192

[8] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed).* Burlington: Morgan Kaufmann, pp. 90-93

[9] Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering* , 1114-1119.

[10] Cutler, A., Cutler, D. R., & Stevens, J. R. (2008). Tree-based Method. In X. Li, & R. Xu, *High-Dimensional Data Analysis in Cancer Research* (pp. 89-109). New York: Springer Science & Business Media.

[11] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed).* Burlington: Morgan Kaufmann, pp. 127-129

[12] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed).* Burlington: Morgan Kaufmann, pp. 124-125