Journal of Korea Multimedia Society Vol. 19, No. 7, July 2016(pp. 1107-1115)
http://dx.doi.org/10.9717/kmms.2016.19.7.1107

A Study on Efficient Market Hypothesis to Predict Exchange Rate Trends Using Sentiment Analysis of Twitter Data

Kokoy Siti Komariah[†], Carmadi Machbub[††], Ary S. Prihatmanto[†††], Bong-Kee Sin[††††]

ABSTRACT

Efficient Market Hypothesis (EMH), states that at any point in time in a liquid market security prices fully reflect all available information. This paper presents a study of proving the hypothesis through daily Twitter sentiments using the hybrid approach of the lexicon-based approach and the naïve Bayes classifier. In this research we analyze the currency exchange rate movement of Indonesia Rupiah vs US dollar as a way of testing the Efficient Market Hypothesis. In order to find a correlation between the prediction sentiments from Twitter data and the actual currency exchange rate trends we collect Twitter data every day and compute the overall sentiment to label them as positive or negative. Experimental results have shown 69% correct prediction of sentiment analysis and 65.7% correlation with positive sentiments. This implies that EMH is semi-strong Efficient Market Hypothesis, and that public information provide by Twitter sentiment correlate with changes in the exchange market trends.

Key words: Efficient Market Hypothesis, Predicting Exchange Rate Trends, Twitter Sentiment Analysis, Hybrid Approach

1. INTRODUCTION

Recent global financial crises have changed international economic relations and have severely affected financial markets. Indonesia Rupiah is no exception, Indonesian currency Rupiah has dropped to its lowest level since 1998 economic crisis, when Indonesia was grappling with the traumatic fallout from the Asian financial crisis. For any country, the foreign exchange rate is considered as a useful barometer for their domestic economy and even a global economy. However, the fluctuation of exchange rate can be influenced by speculations based on information. And in digital era of information, we can access all sorts of information from any sources easily. Today one of the big information sources is social media that include Twitter. Twitter as a social media has a huge implication with rapid and sensitive reaction to political, social, and economic issues. Information has always been important playing a role building market perception and market investment, and Twitter as one of popular social media can be a source of information in predicting currency exchange rate trends based on people's sentiment [1].

The relation between currency exchange rates and sentiment information is believed to be strong but in complicated ways. Important information

※ Corresponding Author : Bong-Kee Sin, Address: (608-737) 45 Yongsro-ro, Nam-Gu, Busan, TEL : +82-51-629-6256, FAX : +82-51-629-6230, E-mail : bkshin@pknu.ac.kr
Receipt date : Feb. 25, 2016, Revision date : June 10, 2016
Approval date : June 15, 2016
[†] Dept. of IT Convergence and Application Engineering, Pukyong National University, Rep. of Korea
(E-mail : kokoy.sitikomariah@gmail.com)
[††] School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia
(E-mail : carmadi@lskk.ee.itb.ac.id)
[†††] Center of ICT Research, Institut Teknologi Bandung, Indonesia
(E-mail : asetijadi@lskk.ee.itb.ac.id)
[††††] Dept. of IT Convergence and Application Engineering, Pukyong National University, Rep. of Korea

frequently results in positive or negative returns. Ryan and Taffler found that for large firms approximately 65% of large price changes and volume movements are linked to publicly available news releases [2] suggests that the movement of exchange rates are affected by those information. And one of financial concept approach developed by Eugene Fama states that at any point in time in a liquid market security prices fully reflect all available information [3].

In this research, in order to prove the hypothesis, we develop a sentiment prediction model from Twitter as a source of information. From a collection of Tweets, we design an efficient prediction model based on machine learning and predict the sentiment in the tweets from tweets timeline. We have chosen IDR or Indonesian Rupiah because of the volatility of IDR Rupiah toward US Dollar. It is reported that IDR Rupiah is changeable and affected by public issues [4]. So, as a way to prove that issues or public opinions can influence the direction of currency movements and market trends, we investigate whether or not Twitter sentiment is correlated with exchange market trends and at the same time prove the efficient market hypothesis (EMH). Of course, true exchange market involves many more factors, but we will just look at the correlation between Twitter sentiments and market trends, merely as an interesting application of sentiment analysis to prove the one of financial concept hypothesis.

2. EFFICIENT MARKET HYPOTHESIS

The efficient market hypothesis (EMH) posits that profiting from predicting price movements is very difficult and unlikely to succeed. The main engine behind price changes is the arrival of new information. A market is said to be "efficient" if prices adjust quickly and on average without any bias to the information and reflect all the available information at any given point in time [3]. Consequently, there is no reason to believe that prices are too high or too low. Security prices adjust themselves before an investor has time to trade on and profit from new pieces of information [2].

The efficient market hypothesis states that market prices mirror all the available information in the market. But the information varies in the degree of the influence security values have. Consequently, financial researchers distinguish three forms of Efficient Markets Hypothesis, depending on what is meant by the phrase "all available information": weak form, semi-strong form, and strong form according to the inclusion of public and private information in the market prices. The weak form of EMH asserts that future prices cannot be fully predicted by analyzing prices from the past. The semi-strong form suggests that current stock prices adapt rapidly to the release of all new public information and the strong form of EMH states that the current prices fully reflect all public and private information.

Papaioannou et al. [5] reported that the information gleaned from microblogging platforms such as Twitter can enhance the forecasting efficiency of Intraday exchange rates. The analysis has shown that efficient market hypothesis supports the analysis of public discussions on Twitter, and vice versa. Seen from the viewpoint of efficient market hypothesis, the analysis of Twitter data helps identifying information which allows us to predict the foreign exchange market.

3. SENTIMENT ANALYSIS TECHNIQUES

Although there are highly accurate methods to analyze and extract relevant knowledge from structured data such as tables or databases, the task of extracting useful information from unstructured data like social media data still remains a major challenge [6]. Sentiment analysis, also known as opinion mining, involves the use of natural language processing, text analysis and compu-

tational linguistics to identify and extract subjective messages. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to a known topic or the overall contextual polarity of being a positive or negative contexts of a document.

There are three sentiment classification approaches can be classified in [7]:

1. Supervised classification, used for predicting the polarity of sentiments based on a training set. Among the machine learning techniques, classifiers based on Naïve Bayes (NB), maximum entropy (ME), and support vector machines (SVM) usually exhibit the best performance [7, 8].
2. Lexicon-based approach. This does not need any prior training in order to mine the data. It uses a predefined list of words, where each word is associated with a specific sentiment.
3. Hybrid approach which combines both machine learning and lexicon based approaches. Mudinas et. al. [9] combined the lexicon based and the machine learning based approaches. In our research we will use this method to extract the Twitter sentiments analysis.

4. PROPOSED METHOD

This research proposes a combination of lexicons and a Naïve Bayes classifier using sentiment words as features about currency exchange rate changes. Fig. 1 shows the organization of the proposed design using Hybrid approach for determining the sentiment polarity from Twitter data.

The framework determines the flow of the research methodology phases which include collection the tweets from Twitter API, text processing, tokenization, feature extraction, and classification. Each step will be detailed in the subsequent sections.

4.1 Pre-Processing

Pre-processing refers to the process of cleaning, normalizing, and tokenizing the data, removing noisy, irrelevant and redundant data for a subsequent classification of the data. The steps of pre-processing are illustrated as follow:

4.1.1 Text Processing

Tweets often contain many words and symbols irrelevant to their sentiments. So, it is desirable to

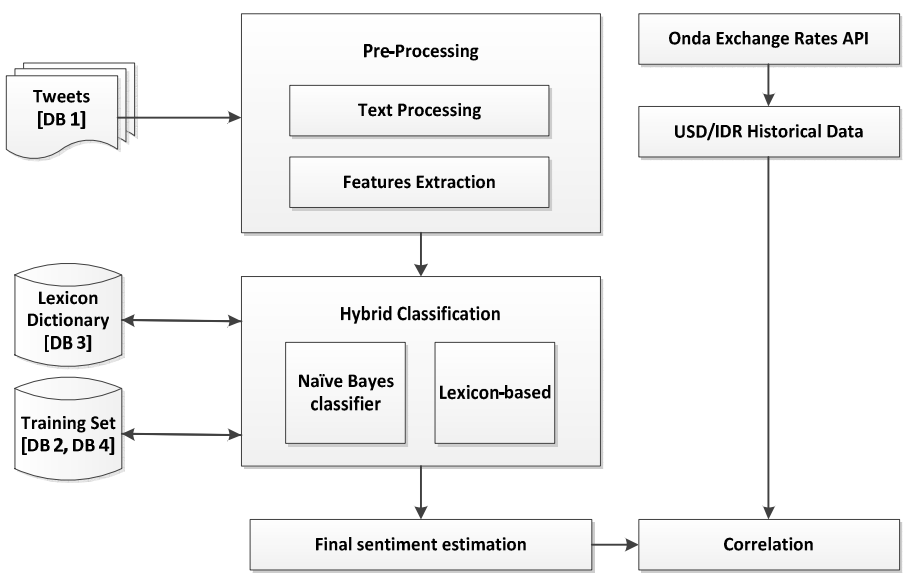


Fig. 1. Proposed research design.

filter them out. We conduct text processing with the following steps:

1. Lower case conversion
   Convert the tweets to lower case.
2. Tokenization
   This involves splitting the text by spaces or delimiters, forming a list of individual words in text. This is also called a bag of words. They will be used as features to train the target classifier.
3. Removing stopwords
   We remove function or stop words from the bag of words by consulting a stopword dictionary. We simply check each word in the bag of words against the dictionary.
4. Twitter symbols
   Many tweets contain special symbols such as "@" for username and or "#" for hashtag, as well as URLs which need to be filtered out entirely as they add no value to the text. To accomplish all of this, we use regex command that searches for matches for these symbols. Additionally, any non-word symbols in the bag of words are filtered out as well.
5. Removing punctuations and additional white spaces. They are simply ignored.

The example of tweet pre-processing step shows as follow:

4.1.2 Feature Extraction

Feature extraction is a very important step in building a successful classifier since a good feature vector largely determines how successful a classifier is. We need a set of feature vectors to build a model of the classifier. In tweets, the presence or the absence of particular words is considered as a feature. The training set consists of positive and negative tweets. Each tweet is split into words and we filter out those words that do not contribute to indicating the sentiment of a tweet. Individual keywords included to the feature vector are referred to as "unigrams". Although not included in this research, we can also consider some other word pair features such as bi-grams and even N-grams. See an example of extracting the tweets as shows in Table 2:

4.2 Classification

In order to determine the sentiment polarity of tweets, this research combines the sentiment lexicons and the Naïve Bayes classifier. For the latter, sentiment words are used as features.

4.2.1 Sentiment Lexicon

A sentiment lexicon, also called a senti-lexicon, is a collection of words which are associated with a polarity score indicating the orientation of the word (positive or negative) and representing intensification or negation. It does not require storing a large data corpus, which makes the whole process much faster. Lexicon construction starts with a small seed set of opinion words which in this re-

Table 1. Tweet Pre-Processing

Before	RT @businessuplift: Since the beginning of the year #rupiah has weakened against #usdollar. http://t.co/PDbF7H3f
After	since beginning year rupiah weakened against usdollar

Table 2. Feature extraction

Tweets	Feature Words
Since beginning year Rupiah weakened against usdollar	"since", "beginning", "year", "rupiah", "weakened", "against", "usdollar"
Rupiah rise due huge investment Korean companies	"rupiah", "rise", "due", "huge", "investment", "korean", "companies"

search came from the AFINN-111 word-list [10] and build an Indonesian version of AFINN-111 to detect Indonesian tweet words. It is followed by a bootstraping cycle to search for synonyms and antonyms in a dictionary corpus iteratively. The following gives an illustration for designing a sentiment lexicon. Table 3 shows the structure of the sentiment lexicon.

In Table 3, the syntactic category refers to the annotated part of speech tag for each word such as adjectives, adverb, nouns, and verbs. The common syntactic category is the adjective. The score refers to the degree of polarity ranging between -5 to 5. This lexicon is stored in separated database.

4.2.2 Naïve Bayes Classifier

Naïve Bayes is a very simple classification model with the ability to classify patterns without using a large training set. This algorithm is very effective in terms of classifying documents [11]. It uses the statistics in order to perform probabilities classification. Basically, it aims to analyze the absence and presence of a particular feature in order to independently classify feature using probabilities. It is very effective when treating words that have probabilities to be opinion or not such as, adjectives or adverbs.

The Naïve Bayes classifier is a probabilistic method that derives from the Bayesian decision theory [12], in Equation 1 the probability of a message d being in class c, $P(c|d)$, is computed by a simple Naïve Bayes model. Although deceptively simple, it has worked well on text categorization [12]. In this model, a given tweet d is assigned the class c^*. It can be formulated as follows:

$$c^* = \arg\max P_{NB}\,(c|d) \tag{1}$$

where Bayesian rule gives the Equation 2:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{2}$$

Given the conditional independent assumption among features f_i in tweet d, we can compute as follows in Equation 3:

$$P(d|c) = \prod_{i=1}^{m} P\,(f_i|c) \tag{3}$$

Then the resulting Naïve Bayes classifier solution is Equation 4:

$$P_{NB}(c|d) = \frac{P(c)\prod_{i=1}^{m} P(f_i|c)^{ni(d)}}{P(d)} \tag{4}$$

Here, f_i represents a feature and $n_i(d)$ is the number of feature f_i found in tweet d, and there are in total m features. Parameters for $P(c)$ and $P(f_i|c)$ are estimated through maximum likelihood estimation along with the add-1 smoothing for unseen features. The reason that this research used Naïve Bayes as classifier is that Naïve Bayes has the ability to classify objects based on its features independently, in other meaning Naïve Bayes focuses on some features regardless of the absence or presence of others for instance, a negative sentiment may consist of negation, adjectives or adverb Naïve Bayes considers each of these features to contribute independently to the probability that the tweet sentiment is negative regardless of the presence or absence of the other features. This scenario is very common in the sentiment analysis where all features may not exist. Therefore, by applying the Naïve Bayes on the tweet data in our research, the classification has been done by identifying the number of positive and negative tweet.

Table 3. Structure of sentiment lexicon

Word	Category	Synonyms	Score	Polarity
Growth	Nouns	Rise, Boost, Build-up	2	Positive
Strong	Adjectives	Powerful, Solid	2	Positive
Fail	Verbs	Breakdown, Fall, Miss	-2	Negative
Weak	Adjectives	Powerless, Weakened, Feeble, Fragile	-2	Negative

5. EXPERIMENTS AND ANALYSIS

5.1 Data Description

The proposed hybrid approach requires an extensive amount of data to construct lexicons and train a classifier. We prepared four databases as described in Table 4. In order to confirm the hypothesis using the hybrid approach, we collected tweets through Twitter API from 6th until 12th August 2015, in total 3235 tweets and stored into DB 1. DB 3 is a collection of positive and negative words and labelled sentences. It contains sentiment words from AFINN word list originally used as features for a lexicon based approach. DB 2 and DB 4 are used to evaluate the Naïve Bayes classifier.

5.2 Sentiments Analysis

We determined the polarity or tendency of the tweets DB 1 either positive or negative. The result is rendered into a bar graph as shown in Fig. 2.

For the first five days, the are more negative tweets than the positive ones, indicating the public opinions or exchange market movement tends to negative, while the results for the remaining two days are indicate the other way round.

In order to evaluate the sentiment-based prediction with respect to the market changes, we plotted two curves as shown in Fig. 3, one for the predicted sentiment analysis trend and the other for the actual exchange rate movement of USD/IDR

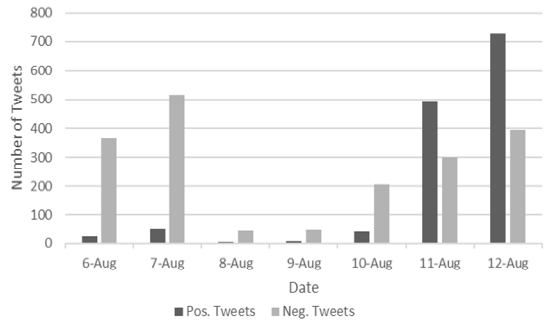


Fig. 2. Daily tweets sentiments distribution.

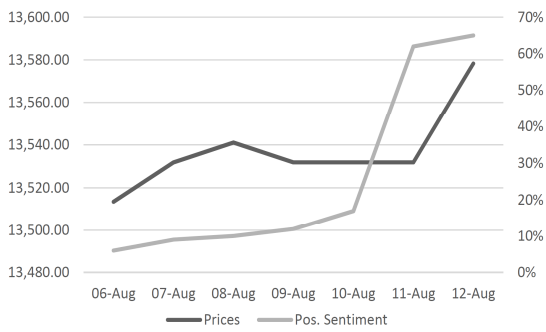


Fig. 3. The comparison of the actual market trends (dark grey) with the Twitter sentiment trends (grey). (Note that sentiments are used to predict market changes)

taken from a financial source [14]. The horizontal axis represents the date of Twitter data while the vertical axis on the left represent the prices IDR Rupiah movement to US Dollar, and the axis on the right represents the volume percentages of positive Twitter sentiment trends.

The grey curve on the graph represents the re-

Table 4. Corpus Data

Corpus	Description	Length
Database DB 1	Twitter dataset collected from 16th until 12th August 2015.	3235 tweets.
Database DB 2	Polarity 2.0 [11].	1000 positive 1000 negative.
Database DB 3	AFINN-111 Word list [10] and Indonesian version of AFINN Words list.	2477 words and phrases with scores ranging from 1 to 5 for positive and -1 to -5 for negative words.
Database DB 4	Training set of tweets from government sites and news accounts related to Rupiah and USD Dollar fluctuation which are labelled as positive and negative.	400 tweets from data corpus which are labelled manually as positive and negative (200 tweets positive and 200 others negative).

sult from prediction model derived from positive sentiment for one-week observation from 6th until 12th August 2015, and the dark curve represent the actual exchange rate (left axis) from financial source [14]. The next two sections we estimate the correlation between the two curves as a way confirming the Efficient Market Hypothesis.

5.3 Classification Result

In order to test the performance of the sentiment prediction algorithm and measure its accuracy we created a confusion matrix. For this we, first, annotated 200 tweets as positive and the other 200 as negative. The test result is summarized in Table 5. There are 48 tweets misclassified into negative class and 76 negative tweets into positive class.

The overall accuracy is 69%, a rate fairly high and deemed enough for practical applications. The next set of tests involves computing the F1 harmonic factor [15] from Table 5 to measure the precision and recall of the sentiment model. It recorded 71% in F-score. In statistical analysis of binary classifications, a high precision means that an algorithm has returned substantially more relevant results than irrelevant ones, while a high recall means that an algorithm has returned many of the relevant results. Based on the confusion matrix in Table 6, calculated all the other measures too, to describe the performance of the classifier and the result shows in Table 6.

5.4 Market Correlation

Given the sentiment polarity of the tweets, it is helpful to calculate the correlation between tweets sentiment and actual currency exchanges from the week-long observations. The two curves in Fig. 3, look similar except for a delay of one day on the 12th day of the actual market curve. But it should be noted that the sentiments are used to predict the market of the following day. In order to prove the assumption regarding EMH, we calculate the Pearson Correlation [16] defined as:

$$r_{xy} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (5)$$

where r is often referred to as the sample correlation coefficient, x_i and y_i are the ith measurements. In this research x is the data from actual market trends and y is the data from the prediction model. The Pearson Correlation between the curves in Fig. 3 is estimated to 65.7%. This confirms the semi-strong form of EMH that posited the current exchange rate of USD/IDR adapts rapidly to the release of all news public information. This result proves that public sentiment information can be a useful indicator and has a strong relation to the fluctuation of currency exchange rate as Fama stated about Efficient Market Hypothesis [3].

Table 5. Confusion matrix evaluation

		Prediction	
		Positive	*Negative*
Annotation	*Positive*	152	48
	Negative	76	124

Table 6. The result test of classification model

Test	Result
Accuracy	69%
Misclassification rate	31%
Recall	76%
Specificity	62%
Precision	67%
Prevalence	50%
F1 harmonic score	71%

6. CONCLUSION

Building a prediction model using hybrid approach combining a carefully designed lexicon and a powerful supervised learning algorithm, this research has proven the efficient market hypothesis with an accuracy of predicted sentiment model of 69%. Pearson Correlation coefficient between the prediction sentiment trends and the actual market trends graph revealed a correlation of 65.7%, which

points to a semi-strong form of efficient market hypothesis. The sentiment scores are shown to be capable of predicting the movement of currency exchange rate quite accurately. The accuracy of prediction model can be further improved by combining other machine learning techniques or by adding different features to provide an even stronger indicator for the currency exchange rate movement. The study of Twitter sentiment analysis in exchange rate trends has shown how efficient market hypothesis can be harnessed to gain insights from data in various social media.

REFERENCE

[1] S.G. Chowdhury, S. Routh, and S. Chakrabarti, "News Analytics and Sentiment Analysis to Predict Stock Price Trends," *International Journal of Computer Science and Information Technologies,* Vol. 5, No. 3, pp. 3595-3604, 2014.

[2] J. Clarke, T. Jandik, and G. Mendelker, "The Efficient Markets Hypothesis," *Expert Financial Planning: Advice from Industry Leaders,* pp. 126-141, 2001.

[3] E.F. Fama, "Efficient Capital Market: a Review of Theory and Empirical Work," *Journal of Finance,* Vol. 25, No. 2, pp. 383-417, 1970.

[4] Indonesia's Weakening Rupiah Challenges Government, Available: http://www.bbc.com/news/business-32357063, (accessed Apr., 19, 2016).

[5] P. Papaioannou, L. Russo, P. George, and C. I. Siettos, "Can Social Microblogging Be Used to Forecast Intraday Exchange Rates?," *Journal Netnomics,* Vol. 14, No. 1-2, pp. 47-68, 2013.

[6] A. Montoyo, P.M. Barco, and A. Balahur, "Subjectivity and Sentiment Analysis: an Overview of the Current State of the Area and Envisaged Developments," *Science Direct: Decision Support Systems,* Vol. 53, No. 4, pp. 675-679, 2010.

[7] M.S. Vohra and J.B. Teraiya, "A Comparative Study of Sentiment Analysis Techniques," *Journal of Information, Knowledge and Research in Computer Engineering,* Vol. 2, No. 2, pp. 313-317, 2012.

[8] S. Kim and B.Y. Hwang, "Prospensity Analysis of Political Attitude of Twitter Users by Extracting Sentiment from Timeline," *Journal of Korea Multimedia Society,* Vol. 17, No. 1, pp. 43-51, 2014.

[9] A. Mudinas, D. Zhang, and M. Levene, "Combining Lexicon and Learning based Approaches for Concept-level Sentiment Analysis," *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '12,* pp. 5:1-5:8, 2012.

[10] F.A. Nielsen, "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs," *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages,* Vol. 718, pp. 93-98, 2011.

[11] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval,* Vol. 2, No. 1-2, pp. 1-135, 2008.

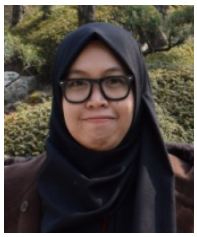
[12] T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[13] C.D. Manning and H. Schütze, *Foundation of Statistical Natural Language Processing*, Massachusetts: MIT Press, Cambridge, 1999.

[14] Oanda, https://www.oanda.com/currency/historical-rates/, (accessed on Mar., 13, 2015).

[15] Wikipedia: F1 Score, https://en.wikipedia.org/wiki/F1_score, (accessed on May, 22, 2016).

[16] Wikipedia: Pearson Product-moment Correlation Coefficient, https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient, (accessed on Mar., 15, 2015).



이Kokoy Siti Komariah

She received her Bachelor degree of Computer Science at Indonesia University of Education in 2011. She later continued her study for a Master degree in Digital Media and Game Technology at Institut Teknologi Bandung in 2014 and joined Master Double Degree Program between Institut Teknologi Bandung and Pukyong National University in 2015. She attended Intelligent Media Lab under supervision of Professor Bong-Kee Sin. Her current research interests are in machine learning, big data and games development.

Carmadi Machbub

In 1991 he got a Doctorat degree in Engineering Sciences majoring in Control Engineering and Industrial Informatics from Ecole Centrale de Nantes, France. As an academic faculty of the Institut Teknologi Bandung, in the period between 1998 until 2011 he was assigned as Head of Electrical Engineering Department, Senior Vice Rector for Resource Management, and Vice Rector for Academic and Student Affairs. He is now Professor and Head of Control and Computer Systems Research Division, School of Electrical Engineering and Informatics, Institut Teknologi Bandung. His current research interests are in pattern recognition, optimization, perception and control.

Ary Setijadi Prihatmanto

He graduated with B.E. and M.S. in Electrical Engineering at Institut Teknologi Bandung in 1995 and 1998, and received his PhD in Applied Informatics from Johannes Kepler University of Linz, Austria in 2006. He is an associate professor & lecturer of School of Electrical Engineering & Informatics, Institut Teknologi Bandung since 1997. He is also the chair of IEEE Computer Society Chapter Indonesia Section and the president of Indonesia Digital Media Forum since 2009. And currently, he has also served as the Head of Information & Communication Technology Research Center at Institut Teknologi Bandung and involved in various governmental, non- governmental and international research projects in Indonesia. His main interests are Human-Content Interaction, Computer Graphics & Mixed-Reality Application, Machine Learning & Intelligent System, Intelligent Robotics & Cyber-Physical System.

Bong-Kee Sin

1985, Bachelor degree from the Department of Mineral and Petroleum Engineering, Seoul National University.
1987, Master degree from the department of Computer Science, Korea Advanced Institute of Science and Technology.
1995, PhD from the Department of Computer Science, KAIST.
1987~1999, Seanior Researcher, SW Research Laboratories, Korea Telecom.
1999~present, Professor in the Department of IT Convergence and Applications Engineering, Pukyong National Universwity.
Research interest : pattern recognition, machine learning, computer vision and artificial intelligence.