Indonesian Journal of Electrical Engineering and Computer Science
Vol. 11, No. 2, August 2018, pp. 791~796
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v11.i2.pp791-796

Opinion Mining using Machine Learning Approach: Case Study of Light Rail Transit Development in Indonesia

Sarifah Putri Raflesia, Dinda Lestarini, Desty Rodiah, Firdaus
Computer Science Faculty, Universitas Sriwijaya
Jl. Palembang- Indralaya KM.33, Indralaya, South Sumatra, Indonesia

Article Info

Article history:

Received Jan 9, 2018
Revised Apr 8, 2018
Accepted May 16, 2018

Keywords:

Opinion mining
Machine Learning
Social Media
Naive Bayes Classifier

ABSTRACT

Light rail transit (LRT), or fast tram is urban public transport using rolling stock similar to a tramway, but operating at a higher capacity, and often on an exclusive right-of-way. Indonesia as one of developing countries has been developed the LRT in two cities of Indonesia, Palembang and Jakarta. There are opinions toward the development of LRT, negative and positive opinions. To reveal the level of LRT development acceptance, this research uses machine learning approach to analyze the data which is gathered through social media. By conducting this paper, the data is modelled and classified in order to analyze the social sentiment towards the LRT development.

Corresponding Author:

Dinda Lestarini,
Computer Science Faculty,
Universitas Sriwijaya,
Jl. Palembang- Indralaya KM.33, Indralaya, South Sumatra. Indonesia.
Email: dinda@unsri.ac.id

1. INTRODUCTION

In Indonesia, the light rail transit (LRT) has been developed in two big cities, Palembang and Jakarta. From the observation through community discussion in real life and social media, the benefit of LRT brings arguments and opinions. Some people claimed that the presence of LRT will bring good value and benefits to citizens and government. Meanwhile, the others agree that the current mass public transportation must be improved.

The opinion, a subjective point of view or judgment for something, have no conclusive statement. But, when opinions come from group of people by means it is generated from social discussion which engaged the stakeholders, the opinions may bring controversy [1].

According to this phenomenon, it is important to do further study, analyze, and measure the level of LRT acceptance in Indonesia. The level of acceptance can be used as one of measurement variables when government needs to analyze and evaluate the LRT development.

In this research, the social media is chosen as field to gather the opinions. Social media contains the social structure such as individual and organization. In social media, people with similar social type are related, they can be families in real life, colleagues, and friends [2]. Social media brought new way in doing interaction, it facilitates people to communicate anytime and anyway without considering how far the distance, time, and places [3].

The social media users often use social media to express themselves by sharing their own information and ideas. It drives the availability of information is limitless which can cause the opinion floods. In this research, social media is used as media to gather the opinion of Indonesian people about LRT. The process of gathering opinion as data to support the analysis is known as opinion mining. It can

be defined as a computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [4]. Moreover, it also aims to determine automatic tool to extract a particular information from natural language text, such as opinions and sentiments. The information will be used to create structured and actionable knowledge to support the decision making process [5]. It is very popular because this approach involves a large amount of data so that the generated information is very objective. In this research, opinion mining was done through the use of social media application programming interface (API).

Machine learning is a well-defined algorithm, data structures and theory of learning, without referring to organism, psychological or evolutionary theory [6]. Machine learning use data to catch a pattern and use the pattern to predict the future data or make a decision in uncertain condition [7]. Data is an example that describe relationship between observed variables. Machine learning use probabilistic theory to build a mathematic model. The mathematic model represents the pattern that explain the relationship between observe variables. The main focus of machine learning research is how to automatically recognize a complex pattern in detail. Eventually, this approach is very helpful to make an intelligent decision based on data. There are several machine learning technique; (1) Support Vector Machine (SVM), (2) k-Nearest Neighbors (KNN), (3) Artificial Neural Network, etc.

In this research, the chosen machine learning technique is Naive Bayes Classifier (NBC). NBC is an algorithm used to find the highest probability to classify testing data into the most appropriate category [8]. NBC can be used in cases that have limited number of target category [9]. It is also known as a simple technique but it has a high accuracy [10] and speed [11]. The advantage of using NBC is it requires a small amount of training data to estimate the parameters necessary for classification [12], [13] so that it has short computational time for training process. NBC simplify learning by assuming that features are independent given class [10]. The classification process is done to the pre-processing data in 2 stages, which are training stage and classification stage. In training stage, training data will be used in learning process to gain knowledge. The second stage is classification stage. In classification stage, system will classify an entity based on the training result. The entity can be classify into positive or negative category. Finally, by conducting this paper, the data is modelled and classified in order to analyze the social sentiment towards the LRT development.

2. RESEARCH METHOD

This part describes about the research methodology to get the research done. Figure 1 shows the research methodology which contains four phases.

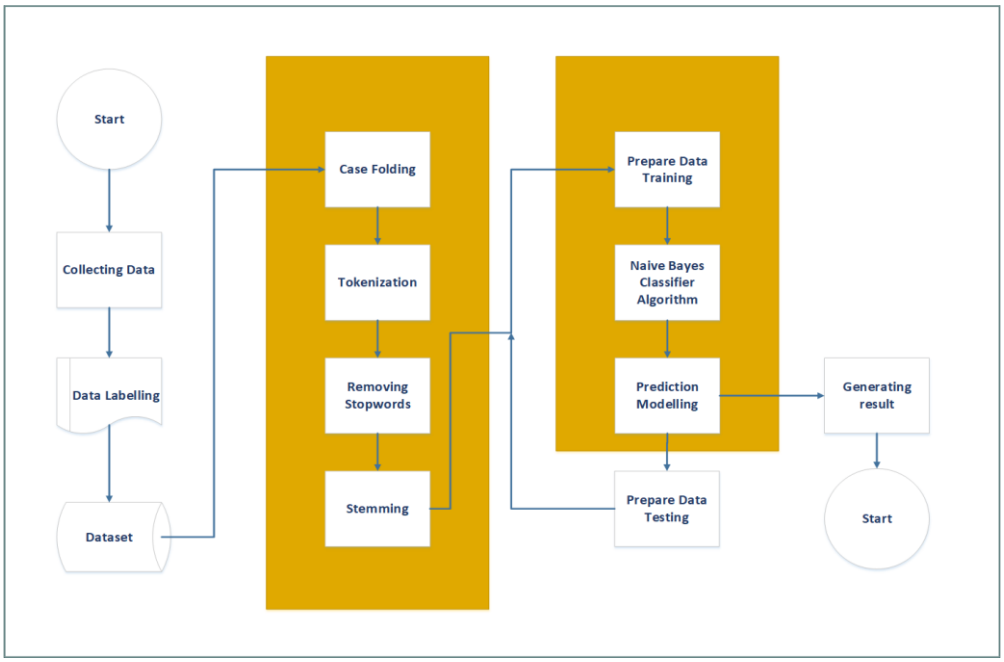

Figure 1. Research Methodology

2.1. Data Collection

The first phase is data collection process. At this phase, people opinion about LRT development were collected from social media. In this research, data were collected from Facebook using Facebook API. This process obtained 494 data from Facebook. Afterward, those opinions are grouped into negative and positive category.

2.2. Pre-processing

The second phase is pre-processing. Pre-processing phase aims to clean up the data from noise. In this research, pre-processing phase contained 4 processes, which are case folding, tokenization, stopwords removal, and stemming.

Figure 2 shows case folding process. In case folding process, the datasets are turned into lower case text. The process is followed by tokenization process as shown in Figure 3. In tokenization process, punctuation marks are discarded and the data are split into a set of words.

```
function __construct($words)
{
    parent::__construct();
    $this->input = strtolower($words);
}
```

Figure 2. Case Folding Process

```
private function splitSentence()
{
    preg_match_all('/\w+/', $this->input, $matches);
    return $matches;
}
```

Figure 3. Tokenization Process

Stopwords removal is used to discard irrelevant words and common words. In this process, a list of common words is created. The list consisted of conjunctions, prepositions or adverbs. The system will compare the datasets and the list of common words. The datasets that contain a word in the list will be removed. Figure 4 shows how this process is done.

```
public function Stopword()
{
    $testData = fopen($testDataLocation, "r");
    $commonWords = array();
    while ($testDatum = fgets($testData))
    {
        $commonWords[] = $testDatum;
    }
    return preg_replace('/\b('.implode('|',$commonWords).')\b/','',$this->input);
}
```

Figure 4. Stopwords Removal Process

The last process in pre-processing phase is stemming as shown in Figure 5. In this process, the words are also reduced by their root word. This is done by removing any attached suffixes and prefixes. A list of suffixes are defined and will be compared with dataset.

2.3. Training

The next phase is training phase. In this stage, training data will be used in learning process to gain knowledge. Naive Bayes Classifiers (NBC) is implemented to calculate the probability of training data. Equation (1) shows the calculation of probability in training data based on Bayes theorem.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

```
protected function stemPluralWord($plural)
{
    preg_match('/^(.*)-(.*)$/', $plural, $words);

    if (!isset($words[1]) || !isset($words[2])) {
        return $plural;
    }

    // malaikat-malaikat-nya -> malaikat malaikat-nya
    $suffix = $words[2];
    if (in_array($suffix, array('ku', 'mu', 'nya', 'lah', 'kah', 'tah', 'pun')) &&
        preg_match('/^(.*)-(.*)$/', $words[1], $words)) {
        $words[2] .= '-' . $suffix;
    }

    // berbalas-balasan -> balas
    $rootWord1 = $this->stemSingularWord($words[1]);
    $rootWord2 = $this->stemSingularWord($words[2]);

    // meniru-nirukan -> tiru
    if (!$this->dictionary->contains($words[2]) && $rootWord2 === $words[2]) {
        $rootWord2 = $this->stemSingularWord('me' . $words[2]);
    }

    if ($rootWord1 == $rootWord2) {
        return $rootWord1;
    } else {
        return $plural;
    }
}
```

Figure 5. Stemming Process

The probability of a category given a document (*P(c|d)*) is calculated by multiplying the probability of a document given a category (*P(d|c)*) and probability of category (*P(c)*) and divided the result with probability of a document (*P(d)*). The probability of a category is simply the number of training documents for a category divided by the total number of training documents. *P(c|d)*, which is also called likelihood, will be used to find data classification.

2.4. Testing

In testing phase, people opinion about Light Rail Transit (LRT) development in Indonesia are collected as data testing. These data will be tested using the probability obtained in the training phase. The probability value is compared with threshold value to determine the classification result. Equation (2) is used to classify data using NBC.

$$C_{NB} = argmax_{c \in C}\ P(c)\ \prod_{x \in X} P(x|c) \quad (2)$$

The opinion can be classify into positive or negative category. If probability value is equal or exceeds threshold value, the opinion will be categorize into positive opinion. On the contrary, the opinion can be categorize into negative opinion when the probability is below threshold value.

3. RESULT AND ANALYSIS

The implementation of classification process is conducted using phyton and PHP. Evaluation process aims to measure performance of NBC in classifying an opinion in to its respective class. The amount of data that we used as testing data in evaluation process were 126. Table 1 shows confusion matrix for classification process.

Table 1. Confusion Matrix for Naive Bayes Classification

Data Class		Prediction		
		Positive	Negative	Total
Actual	Positive	65	4	69
	Negative	17	40	57
	Total	82	44	126

Data in confusion matrix will be used to evaluate performance of classification process. The common measurements for classification based on the confusion matrix [14] are shown on Table 2.

Table 2. The Measurement of Performance

Measurement	Value
Precision	0.79
Recall	0.94
Specificity	0.70
F-Measure	0.86
Accuracy	0.83
Area Under the Curve (AUC)	0.82

Precision describes the proportion of pedicted positive cases that are correctly classified [15]. In this research, the precision value of classification process is 0,79. This result is quite high which indicates that the classification result only have small number of false positive. Sensitivity and specificity are two measurements that are used together to measure the predictive performance of a classification model [16]. Sensitivity, which is also known as recall, shows the proportion of real positive cases that correctly classsified. Specificity describes the proportion of real negative cases that are correctly classified. The recall value of classification process is 0,94. This result shows that the classification process produces a very small amount of false negative. The result also shows that the specificity value in this research is not quite high. System can only classified 70% real negative cases correctly.

In addition to those measurements, we also use F-measure to evaluate classsification result. F-measure is considered as a better measurement than precision and recall because it takes both precision and recall measurement into consideration. F-measure will produce a high result when precision and recall value are balance. The value of F-measure in this research is 0,86.

Accuracy is a measurement to evaluate ratio of correct prediction cases over the total number of cases [17]. Overall accuracy of classification using naive bayes is 0.83. This result shows that the system can correctly classified 83% cases over all the given cases.

AUC is a measurement to evaluate the ability of classifier in avoiding false classification [14]. AUC is believed as a better measurement to evaluate a machine learning technique than accuracy because AUC is more discriminating and statistically consistent [18]. As shown in Table 2, AUC value in this research is 0.82.

4. CONCLUSION

In this research, we used machine learning approach to classify user perception of Light Rail Transit in Indonesia. We implemented NBC to determine probability and likelihood ratio. Naive bayes classifier had been chosen because it requires a small amount of training data and short computational

time for training process. Datasets were collected from Facebook using Facebook API. The testing result shows that the technique is quite effective in classifying people opinion about LRT development. This result also indicates that the technique can be used to gain knowledge in order to support decision-making process regarding LRT development in Indonesia.

REFERENCES

[1] N. Anstead and B. O'Loughlin, "Social media analysis and public opinion: The 2010 UK general election," *J. Comput. Commun.*, vol. 20, no. 2, pp. 204–220, 2015.
[2] J. A. Barnes, "Graph theory and social networks: A technical comment on connectedness and connectivity," *Sociology*, vol. 3, no. 2, pp. 215–232, 1969.
[3] A. Whiting and D. Williams, "Why people use social media: a uses and gratifications approach," *Qual. Mark. Res. An Int. J.*, vol. 16, no. 4, pp. 362–369, 2013.
[4] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, Springer, 2012, pp. 415–463.
[5] E. Fersini, E. Messina, and F. A. Pozzi, "Sentiment analysis: Bayesian ensemble learning," *Decis. Support Syst.*, vol. 68, pp. 26–38, 2014.
[6] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Mach. Learn.*, vol. 3, no. 2, pp. 95–99, 1988.
[7] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. 2012.
[8] R. Feldman and J. Sanger, "The text mining handbook: advanced approaches in analyzing unstructured data," *Imagine*, vol. 34, p. 410, 2007.
[9] Z. F. Alfikri and A. Purwarianti, "Detailed Analysis of Extrinsic Plagiarism Detection System Using Machine Learning Approach (Naive Bayes and SVM)," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 12, no. 11, pp. 7884–7894, 2014.
[10] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41–46.
[11] L. Duan, P. Di, and A. Li, "A New Naive Bayes Text Classification Algorithm," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 12, no. 2, pp. 947–952, 2014.
[12] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier," *arXiv Prepr. arXiv1610.09982*, 2016.
[13] L. Fan, X. Huang, and L. Yi, "Fault Diagnosis for Fuel Cell Based on Naive Bayesian Classification," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 11, no. 12, pp. 7664–7670, 2013.
[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.
[15] D. M. W. POWERS, "Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[16] K. M. Ting, "Sensitivity and Specificity," in *Encyclopedia of Machine Learning*, C. Sammut and G. . Webb, Eds. Boston, MA: Springer, 2011.
[17] M. Hossin and M. N. Sulaiman, "a Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 1–11, 2015.
[18] C. X. Ling, J. Huang, and H. Zhang, "AUC: A statistically consistent and more discriminating measure than accuracy," in *IJCAI International Joint Conference on Artificial Intelligence*, 2003, pp. 519–524.