Simulation of Marketplace Customer Satisfaction Analysis Based on Machine Learning Algorithms

Ajeng Aulia Turdjai
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Ganesha Street 10, Bandung 40132, Indonesia
auliajeng@students.itb.ac.id

Kusprasapta Mutijarsa
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Ganesha Street 10, Bandung 40132, Indonesia
kusprasapta.mutijarsa@ itb.ac.id

***Abstract*— Twitter can be a source of public opinion data and sentiment. Such data can be used efficiently for marketing or social studies. In this research addresses this issue by measuring net sentiment based on customer satisfaction through customer's sentiment analysis from Twitter data. Sample model is built and extracted from more than 3.000 raw Twitter messages data from March to April 2016 of top marketplace in Indonesia. We compared several algorithms, and the classification schemes. The sentiments are classified and compared using five different algorithms classification. There are, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine. The five machine learning can be applied to the Indonesian-language sentiment analysis. Preprocessing on the stages of tokenization, parsing, and stop word deletion of word frequency counting. Frequency of the word of the document used weighted by TF-IDF. The Random Forest, Support Vector Machine, and Logistic Regression generate better accuracy and stable compared with the Naïve Bayes and K-Nearest Neighbor. The results showed Support Vector Machine has accuracy 81.82% with 1000 sampling dataset and 85.4% with 2000 sampling dataset. This shows that the more the number of training data will improve the accuracy of the system. The Net Sentiment score for marketplace in Indonesia is 73%. This results also showed that customer satisfaction has average Net Promoter Score (NPS) 3.3%.**

Keywords— Marketplace, Customer Satisfaction, Machine Learning, Algorithms Clasification

I. INTRODUCTION

Based on data from the Socio Economic Status (SES) in 2014, there are 92 million or more than 40% of the bank account linked to a credit card and a debit of Indonesia's population reached 240 million. When compared with the penetration of the mobile phone, the rate is still low because about 85% of Indonesia has a mobile phone where each month they spend 661 pages for browsing [1].

In 2015, We Are Social, a social marketing agency reported on the development of e-commerce based devices used in Indonesia society. There are 18% of active users who accessed the internet through a computer to browse products that may be purchased, and 16% of them purchase goods online at one of the sites accessible. While active users who access the internet via mobile phones have 11% to see the products that may be purchased, while 9% are buying goods online that are accessed through a mobile phone [2].

At this time microblogging site has become a very popular means of communication among internet users. Where millions of messages every day on a popular website that provides microblogging services such as twitter, tumblr, and facebook [3]. This led to more and more users are posting about product and service they use, or express their views on polotics and religion. In 2014, Indonesia has 20 million active users twitter [4]. Twitter can be a source of public opinion data and sentiment. Such data can be used efficiently for marketing or social studies [5].

Twitter sentiment analysis on weaknesses in the words contained in sentences posted by users of the site. Twitter only allows users to write 140 characters, which is why users often use abbreviations and spelling words wrong word. The wrong way of writing resulted in deficiencies in the process of text mining, which can complicate the features taken as well as reduce the accuracy of the classification.

The magnitude of the effect and benefits of Sentiment Analysis, leading research or application of the Sentiment Analysis is growing rapidly, even in America, there are approximately 20-30 companies use Sentiment Analysis to obtain information about public sentiment towards the company's service [6].

Sentiment analysis have influence and benefits, which is to obtain information about public sentiment towards companies [7]. As previous research associated with Sentiment Analysis, among others, research [8] detects fake websites or sites with the original classification of news articles on the website. Research [9] to analyze sentiment on twitter text, using the n-gram language characters and SVM models to cope with high lexical variation in Twitter text. Research [10] developed a system that can identify and classify public sentiment to predict interesting products in marketing.

Driven by this phenomenon, in addition to marketing with advertising on television, many e-commerce companies are trying to gain more customers with social media marketing. This research examines based on the analysis of customer satisfaction of online shopping based marketplace based on sentiment. Marketplace in terms of this research is the online market. Marketplace companies do not sell their own products, they do not have to stock the product, they only act as a mediator or parties are providing facilities and systems so that transactions between buyers and sellers can easily be done. Data tweets were taken from the marketplace in Indonesia frequented its website by internet users [11].

Automated sentiment analysis focuses on analyzing the content of online posts, determining whether they are positive, negative, or neutral, and aggregating the sentiments detected into a single generic score. The Net Sentiment Score computes a ratio of positive and negative mentions of a topic. Currently, many companies uses NPS (Net Promotor Score) to measure their customer loyalty and satisfaction [12].

Therefore, this research will examine the net sentiment analysis based on sentiment of tweets that are in Indonesia which will then be classified using a K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest. This research is expected to accelerate efforts to obtain accurate information about the sentiment of the customer marketplace on a matter.

II. THEORITICAL BASES

A. Text Mining

Text mining is one of the applications of data mining. Text mining is one form of text data exploration and analysis that aims to gain new knowledge either through automated or semi-automated [13]. Text mining can be defined as a process of digging up information which a user interacts with a set of documents using tools to analyze documents that are components of the data mining that dalah them is categorization [13]. The purpose of text mining is to obtain useful information from a collection of documents. Source of data used in text mining is a collection of texts that have a format that is not structured.

Text mining refers to the process of taking high-quality information from text. High quality information is usually obtained through forecasting patterns and trends through means such as learning statistical patterns. Typical text mining process include text categorization, text clustering, extraction concept / entity, sentiment analysis, inference documents, and entity relationship modeling.

B. Sentiment Analysis

Analysis sentiment also called opinion mining, is one branch of science of data mining that aims to analyze, understand, process, and extract the textual data of opinion against entities such as products, services, organizations, individuals, and specific topics [14]. This analysis is used to obtain certain information from an existing data set. Sentiment analysis focuses on the processing of opinions containing polarity, which has a value of positive or negative sentiment.

C. Machine Learning

Machine learning is a discipline that gives a computer the ability to learn without explicitly in the program. In machine learning algorithms can be grouped based on expected input and output of the algorithm.

1. Supervised learning, create a function which maps input to output desired. Such as grouping (classification). A learning algorithm is based on a sample set of input-output pairs are desirable in large enough quantities. This algorithm observe these examples and then produce a model capable of mapping the new input into the appropriate output [15].
2. Unsupervised learning, modeling the set of inputs, such as classification (clustering). This algorithm has the objective to study and look for interesting patterns on a given input [16]. Although not provided proper output explicitly. One of unsupervised learning algorithms most commonly used is clustering / grouping [15].

III. DESIGN

A. Flow Classification of Sentiment

Sentiment analysis system developed as a system that can classify the type opinion into positive, negative, and neutral by using five methods. The methods used for different machine learning classifiers, consisting of K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest.

The data retrieved is data tweet containing sentiment towards the marketplace (online shopping). In this system, the raw data obtained in advance separated into two types of data. They are training and testing data.

Fig. 1. Sharing of Raw Data

In the process of training and testing data are crawling data. Stages of the process of training data is done by crawling the data, then straight to the preprocessing (tokenizing, stemming, stop word, and filtration), then give judgment in the form of weighting word, and the learning process with five methods of Machine Learning. Sentiment classification is done at this stage, the results of the classification can result in the evaluation or accuracy of the results obtained by training data on these five methods.

Training data is used as input into the system consists of a collection of data tweet, already labeled type of sentiments, such as positive, negative, or neutral. The results of the training process is a model of sentiment analysis in the form of a probabilistic model every word to every type of sentiment.

While the test of data used as input into the system consists of a data set to be tested tweet. The results of the test process is a collection of tweets that have been classified into types

sentiment is positive, negative, or neutral. The data is immediately processed through a learning process with five methods of Machine Learning or referred to sentiment classification. Flow of the testing process can be seen in Fig. 2.

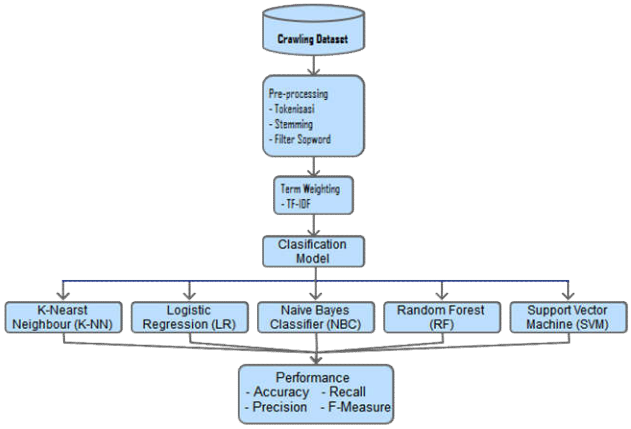

Fig. 2. Flowchart Classification

B. Preprocess

The raw data in the form of a text obtained from social networks usually have a high noise level and structure of language which slightly irregular. On the social networking twitter, non-standard use the word so widely used that it takes some preprocessing stages is done so that data becomes more simple and structured so that it can be easily processed to the next stage. Steps being taken on the document preprocessing is cleaning, to reduce noise when analyzing sentiment. Preprocessing steps are performed in this research are:

Fig. 4. The workflow of our sentiment analysis

- Conducting the process of folding case, that is by any part of the tweet converted into written with lowercase letters.
- Tokenization is cutting words in each sentence. This is done so that each word can be seen at a frequency of occurrence of the news. It works on the word weighting. At this stage of tokenization do remove punctuation. Tweets that have been selected, the tweet is divided into sections separated by a space character.
- Doing normalization features, which throw out some typical components commonly found in the tweet, such as the URL address, and retweet (RT).
- Removing stopwords, at this stage to omit important not appropriate data dictionary used, in order to increase the accuracy of the weighting term.
- Stemming, in this research using Indonesian stemming from Tala who concluded a basic word can be added affixes [7].

C. Weighting Term

Term weighting matrix is a document that represents a collection of documents used to process text document classification. This research will be used as a method of TF-IDF weighting process, which will be the weighting of each term based on the level of interest in a set of input documents.

D. Model Sentiment Analysis

In the modeling and classification step, in this research use five classifiers algorithm: K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest. About more than three thousands of tweets have been used to build the training set.

In K-Nearest Neighbor classification process each document will be positive or negative sign, if the number of positive < number of negative then a score of sentiment:

$$-1\ x\ \left(\frac{number\ of\ negative\ word}{number\ of\ words}\right) \quad \ldots(1)$$

If, number of positive > number of negative then a score of sentiment:

$$\left(\frac{number\ of\ positive\ word}{number\ of\ words}\right) \quad \ldots(2)$$

If in addition to the above criteria, then the sentiment is 0 or a so-called neutral. Then it will do the process K-Nearest Neighbor to calculate value of the similarity between test data with all training data on the document by using the cosine similarity. Fig. 5 shows process workflow of K-Nearest Neighbor classification.

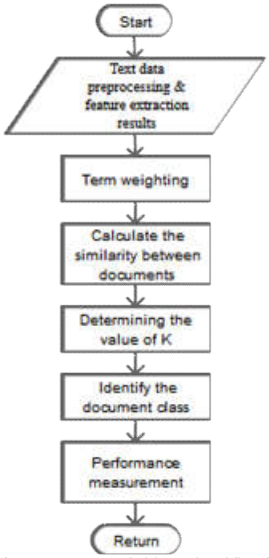

Fig. 5. The workflow of K-Nearest Neighbor classification.

Logistic Regression is statistical classification model. Predicts binary response from a binary predictor for predicting the outcome of a categorical dependent variable. Logistic regression measures the relationship between a categorical dependent variable and one or more independent variable.

Logistic Regression classification Logistic regression was divided into two types, they are binomial or binary logistic

regression and multinomial logistic regression. Binomial deals with variable in which the observed outcome have two possible types in this cases are positive or negative. Outcome is coded as 1 or 0. Binomial is straight forward interpretation. Multinomial logistic regression deals with situation where there are three or more outcomes.

Naïve bayes classification from bayes rule, with the basic formula: $P(Y|X) = \frac{P(X)P(Y|X)}{P(Y)}$ …(3)

From the formula can be made an assumption about how we calculate the probability of the emergence of the document which is equivalent to multiplication (product) on the probability of occurrence of each word in it. This causes no relationship between one word with another word.

From the bayes rule can estimate the probability of occurrence of a word as a positive or negative sentiment to view the training data set of positive and negative sentiment and counting how often words that appear in each class. This is what makes this training as guided learning. Stages of the classification process with NBC shown in the workflow as in Fig. 6.

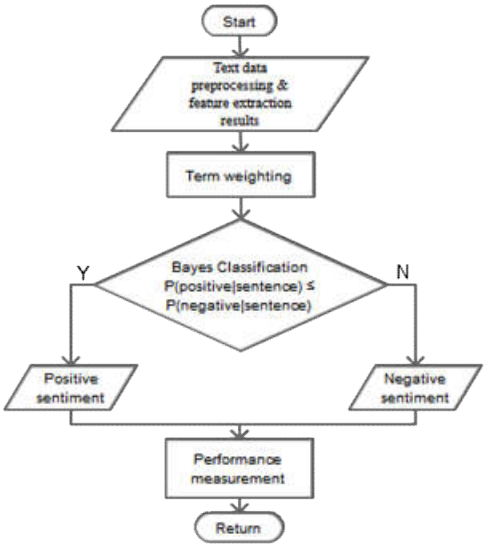

Fig. 6. The workflow of Naïve Bayes classification.

Random forest is a collection of some trees, each of the tree depends on the pixel values of each vector are taken randomly and independently. Random forest is not inclined to overfit and can make the process fast, making it possible to process as many trees as desired by the user.

Fig. 7 shows the workflow random forest algorithms. In the establishment of tree, random forest algorithms will be doing training from sampling data. After all the tree is formed, then the classification process will run. Class determination done by voting on each tree, the class with the highest number of votes will be the winner.

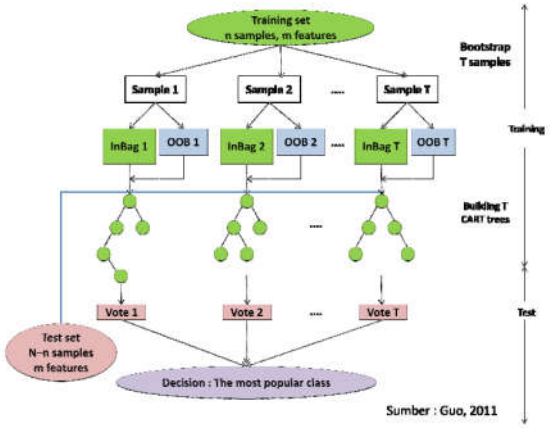

Fig. 7. The workflow of Random Forest classification.

Support Vector Machine classification process uses by converting text into vector data. Vectors in this research had two components, the dimension (word id) and weights. The weights are often combined into a tf-idf value, simply by multiplying them together. Process flow diagram of classification with Support Vector Machine is shown in Fig. 8.

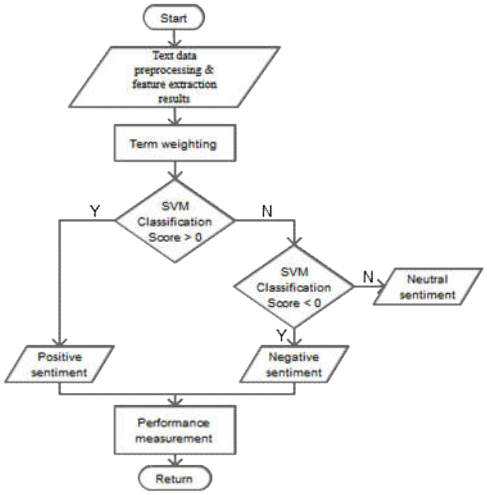

Fig. 8. The workflow of Support Vector Machine classification.

IV. Experiment Results And Analisys

Processed dataset conducted by uniformity standard words from the dictionary. The document has been done in the preprocessing stage then analyzed by using machine learning. In this research used the comparison between the five-supervised learning, that are: K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest.

A. Analysis Net Sentiment and Net Promotor Score

In the testing stage of sampling data to perform feature dataset positive, negative, and neutral is done by crawling text data document using crawler4j. The text file with each record contains a sentence with positive sentiments, negative, and neutral. In this research the dataset used is related to the existing marketplace in Indonesia, crawling outcome document

consists of a 3150 tweet. The results of the test data crawling positive, negative, and neutral is not the same as crawling data is not selected manually. Data crawling results directly in the process and the system will select into 3 categories: data, including positive, negative, and neutral. Results from crawling tweet can be seen in the Table 1.

Table 1. Total crawling dataset

Crawling Dataset	Total Tweet	%
Positive	529	16.7936508
Negative	423	13.4285714
Neutral	2198	69.7777778
Total	3150	100

From Table 1 we can calculate the Net sentiment (sentiment net value), that's the opinion or feelings of consumers expressed in social media to a brand in the world of social media. Results of net sentiment and net promoter score can be seen in the Figure 3. In this research resulted in net sentiment score of 11%.

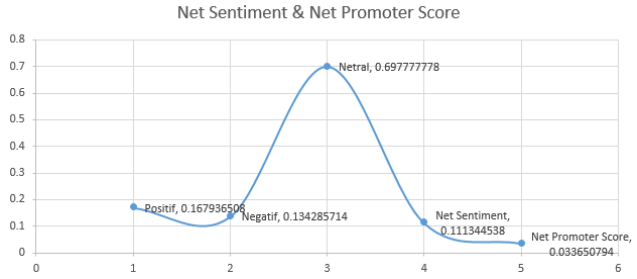

Fig. 9. Net sentiment and Net Promoter Score customer marketplace

From Figure 9, NPS shows the percentage of the value of customer satisfaction of customer sentiment marketplace in Indonesia at a certain time. In this research the value of NPS obtained is still very small, NPS values obtained for 3.36%. If the value of NPS reaches 100%, meaning that all customers are promoters. Based Satmetrix survey in 2013 [17], Apple (Computer Hardware) has a value of NPS of 72%, Google (search engine) 53%, Amazon (online shopping site) 69%.

B. Analysis and Evaluation Algorithms

Dataset processed in this research only the positive and negative sentiment. The examination based on classification algorithm using cross validation technique. On machine learning we use 10-folds cross validation by library sklearn. The results of extensive experiments and theoretical evidence, showed that the 10-folds cross-validation is the best choice to get an accurate validation results. 10-fold cross-validation testing will repeat 10 times and the measurement result is the average value of 10 times the test.

In this research used compare of five machine learning classification algorithms, consist of K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest.

All machine learning will perform three times of experiment. In the first experiment of 1000 data is divided into two sections of data, training data and testing data, the ratio of training data: testing data = 70%:30%. The second experiment of 1000 the same data from the first experiment, the ratio of training data: testing data = 60%:40%. A third experiment to 2000 data by comparing training data: testing data = 60%:40%. Comparison test result of each machine learning can be seen in Table 2.

Table 2. Experiment Results to search The Best machine learning.

Experiment	KNN	LR	NB	RF	SVM
1	0.667	0.798	0.75	0.775	0.818
2	0.65	0.788	0.73	0.767	0.808
3	0.6125	0.829	0.7625	0.8	0.854

The experiment on algorithm showed that the best classification algorithm is Support Vector Machine algorithm with the highest score. From the results that have been tested can be viewed in detail in Table 2. The results showed K-Nearest Neighbor has the smallest accuracy than the five machine learning. Three machine learning which has the highest accuracy results obtained by Support Vector Machine, then Random Forest, and Logistic Regression.

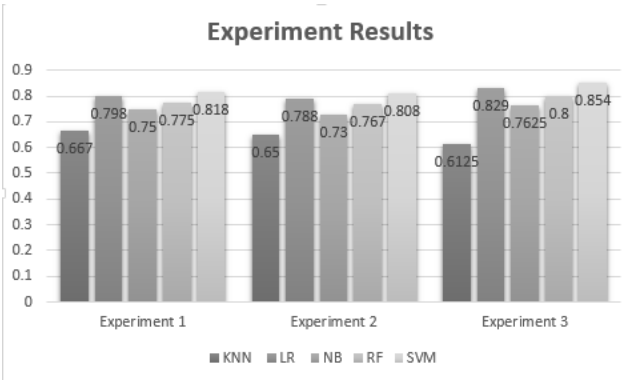

Fig. 10. Experiment Results accuracy of K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest.

Table 3. Performance classification machine learning algorithms in the 3rd experiment.

Algorithm	Accuracy	Precision	Recall	F-Measure
KNN	0.6125	0.67	0.61	0.58
LR	0.829	0.8	0.8	0.8
NB	0.7625	0.76	0.76	0.76
RF	0.8	0.82	0.8	0.8
SVM	0.854	0.87	0.85	0.85

From Table 2 shows based on the calculation of the percentage of accuracy, the accuracy of the highest value obtained in the third experiment at 85.4% were performed using SVM method with a dataset of 2,000 data. In this research, the topic of customer sentiment in Indonesia machine learning marketplace nicest and most stable accuracy is SVM.

Support Vector Machine provides a good process because of the pattern of Support Vector Machine method refers to the availability of support vector to form hyperlane. The concept of SVM can be explained simply as an attempt to find the best hyperplane which serves as a separator are two classes in the

input space. Efforts to locate the optimal hyperplane is the core of the learning process on SVM. Hyperplane best separation between the two classes can be found by measuring margin or the distance between hyperplane the data closest from each class hyperplane and seek the maximum extent.

The data in this research tended distributed linear SVM which has the advantage of analyzing the data that is distributed linearly. Then the number of test data affects the outcome of generalizations. But the results will be better if the number of learning data more or equal to the number of test data. From the results of the computer simulation shows that Support Vector Machine provides good results in predicting test data.

CONCLUSION

The results showed the method K-Nearest Neighbor (K-NN), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) can be applied to the Indonesian-language sentiment analysis.

Support Vector Machine, and Logistic Regression generate better accuracy and stable compared with the Random Forest, Naïve Bayes and K-Nearest Neighbor. Support Vector Machine better than Logistic Regression and the other machine learning. Support Vector Machine provides a good process because of the pattern of Support Vector Machine method refers to the availability of support vector to form hyperlane. In this research Support Vector Machine has advantages in analyzing the data that is distributed linearly. Support Vector Machine has an average degree of accuracy.

The results showed Support Vector Machine has accuracy 80.8% with 1000 sampling dataset and 85.4% with 2000 sampling dataset. Logistic Regression has accuracy 78.8% with 1000 sampling dataset and 82.9% with 2000 sampling dataset. This shows that the more the number of training data will improve the accuracy of the system.

Net Sentiment Scores derived from social media opinions. In this research also showed that Net Sentiment score 11%. This results also showed that customer satisfaction has average Net Promoter Score (NPS) 3.3%. This results showed that the majority of online shopping customers were less satisfied in shopping online in the marketplace in Indonesia.

REFERENCES

[1] Anonymous, "Socio Economic Status (SES) Jakarta," 25 July 2015. [Online]. Available: http://belanjanesia.com/news/4/Perilaku-konsumen-di-Indonesia-terhadap-belanja-online.html. [Accessed 11 November 2015].

[2] S. Kemp, "DIGITAL, SOCIAL & MOBILE IN 2015," We Are Social, 21 January 2015. [Online]. Available: http://wearesocial.com/sg/special-reports/digital-social-mobile-2015. [Accessed 21 November 2015].

[3] Anonymous, "The top 500 sites on the web," Alexa Internet, 2016. [Online]. Available: http://www.alexa.com/topsites. [Accessed 5 February 2016].

[4] Anonymous, "Asia Pacific: Twitter visitor reach 2015," Statista, 2016. [Online]. Available: http://www.statista.com/statistics/254803/twitter - penetration-in-selected-apac-countries/. [Accessed 25 November 2015].

[5] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *In Proceedings of LREC,* 2010.

[6] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 417-424, 2002.

[7] F. Z. Tala, "A Study of Stemming Effects on Information," Universiteit van Amsterdam, The Netherlands, 2003.

[8] A. Abbasi, Z. Zhang and C. Hsinchun, "A Statistical Learning Based System for Fake Website Detection," *The Workshop on Secure Knowledge Management,* 2008.

[9] Q. Han, J. Guo and H. Sch ütze, "Combining an SVM Classifier and Character N-gram Language. Models for Sentiment Analysis on Twitter Text," *Association for Computational Linguistics: Human Language Technologies,* 2013.

[10] V. G and R. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, no. 6, pp. 283-293, 2012.

[11] Anonymous, "Top sites in Indonesia," Alexa Internet, 2016. [Online]. Available: http://www.alexa.com/topsites/countries/ID. [Accessed 5 February 2016].

[12] N. A. Vidya, M. I. Fanany and I. Budi, "Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers," in *The Third Information Systems International Conference*, Surabaya, 2015.

[13] R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, 1st Edition ed., New York: Cambridge University Press, 2007.

[14] R. Moraes, J. Valiati and W. P. Gavião Neto, "Document-level sentiment classification: an empirical comparison between SVM and ANN," *Expert Systems with Applications: An International Journal,* vol. 40, no. 2, pp. 621-633, 2013.

[15] S. Jonathan, R. P. Norvig, J. F. Canny, J. M. Malik and D. D. Edwards, Artificial Intelligence: A Modern Approach, vol. vol. 2, New Jersey: Prentice Hall, Englewood Cliffs, 1995.

[16] K. P. Murphy, Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series), 1st Edition ed., The MIT Press, 2012.

[17] L. P. Rikasasi , "Ketahui Tingkat Kepuasan Pelanggan Anda dengan Metode NPS," 28 January 2013. [Online]. Available: http://mebiso.com/menghitung-tingkat-kepuasan-pelanggan-melalui-nps/. [Accessed 25 March 2016].