Application of Sentiment Analysis on Product Review E-Commerce

Yuniarta Basani, Harris V. Sibuea, Sinta Ida Patona Sianipar and Jen Presly Samosir
Informatics Technology, Institut Teknologi Del, Laguboti 22381, Indonesia

Abstract. The lack of buying product via online are the consumer cannot touch, try, or even see it directly. Then how does the consumer believe the product they like, is the correct product to be bought. The main key is product review from the consumer who have bought and try the product. The more number of product review on certain product or the popular product caused difficulties for the consumer to decide which product they should choose. For that we require a solution, which is the writer will apply sentiment analysis to classify every product review into positive orientation, negative or neutral and also produce summary of product review based on product feature to help reading process, product review and decision making. The writer step is (1) data collecting and preprocessing, (2) product feature extraction using Double Propagation, (3) deciding sentiment orientation, (4) classifying by using Naïve Bayes Classifier and Support Vector Machine, and (5) summary generation based on product feature. These stages run in a simulator and the result of the classification from both methods compared.

1. Introduction
The number of internet users and e-commerce sales in Indonesia continues to increase. The rapid expansion of the use of e-commerce makes more and more people buying products online. E-commerce has great potential, but there are some problems that hinder growth or flaws facing e-commerce. The disadvantage of buying products online is that consumers can not touch, try or even look directly at the product. Then how consumers can believe that the product they like, is the right product to buy. The main key is product reviews from consumers who have bought and tried the product (Hu & Liu, 2004). It becomes the basis that it should be given an ease of access, both for consumers, ecommerce owners and also product manufacturers to view reviews easily and appropriately (Hu & Liu, 2004). Therefore, it is necessary to apply sentiment analysis to label every product review based on customer's review text in product review which they write, positive, negative or neutral and produce summary of product review based on product feature. Brief introduction and product review classification based on product features for consumers will be an attraction or added value in some sites such as www.rottentomatoes.com, a trusted site in the quality assessment of a movie or TV show based on hundreds of reviews covering movies or TV show (Anon., 2010), where all the reviews are briefly displayed and calculate the percentage of consumers who rated the film or TV show as positive or not (Pang, et al., 2002).

Related research is the research in 2004 conducted by Hu and Liu. The research applies sentiment analysis on product review of 5 types of electronic products: 2 types of digital cameras, 1 type of mp3 player, and 1 type of cellular phone. The product review data is collected from the Amazon.com...
dataset. Hu and Liu also produced a product review overview based on the features of the product. These summaries make it easy for potential customers to feel the same way about products that have been purchased by previous customers. Potential consumers who are very interested in certain features can easily browse whether consumers who have purchased a product feel satisfied or complain about the product.

Research on this final task is to determine the classification based on sentiment orientation. The author uses the classification method of Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) in determining the positive, negative, or neutral label of a product review. In the final stage, the authors produce a summary of the classification results grouped by features of each product and compare the two methods.

2. Supervised Learning Method

Supervised Learning is a supervised learning where if the expected output has been known before. Usually this learning is done by using existing data. Supervised learning is a method used to find relationships between input attributes (can be called as independent variables) and target attributes (can be referred to as variable dependent). The relationship is found as a representation of a structure called a model.

Classification is one of the data mining techniques used to predict a class of existing data. This classification is done because there are a number of rows of data that are known to be class and labeled. Classification consists of two steps, namely the development of models of training data and the use of models or application of models that have been built from training data to data testing. Some methods that can be used in the classification process, including the Support Vector Machine, Naive Bayes Classifier, Decision Tree, Random Forest and so on. The method used is the Naive Bayes Classifier and Support Vector Machine.

2.1. Naïve Bayes Classifier

A Naïve Bayes Classifier (NBC) is a simple probabilistic classifier based on the Bayesian theorem's application (with Bayesian statistics) with strong independent (naïve) assumptions. A more descriptive term for the underlined probability model is the “independent feature model”.

Depending on the exact situation of the probability model, NBC can be trained very efficiently in supervised learning. In practical applications, the estimation parameters for the NBC model use the maximum likelihood method; in other words, one can work with the Naïve Bayes model without trusting Bayesian probability or using other Bayesian methods. The classification using the Naïve Bayes algorithm is based on the Bayes theorem (Han et.al, 2012). Naïve Bayes Classifier (NBC) has the formula (1).

\[
(C_i | X) = \frac{(X|C_i)P(C_i))}{(P(X))}
\]

with

\[X = \text{data, } x_1, x_2, \ldots, x_n\]
\[C_i = \text{class, } C_1, C_2, \ldots, C_m\]
\[P(X) = \text{probability of data } X\]
\[P(C_i) = \text{class probability } i\]
\[P(X | C_i) = \text{probability of data } X \text{ in class } C_i\]
\[P(C_i | X) = \text{probability of class } C_i \text{ on } X \text{ data.}\]

To determine which data belong to which class, is determined by obtaining the probability of the largest class of the same data (X), expressed in formula (2)

\[P(C_i | X) > (C_j | X) \text{ for } 1 = j = m, j \neq i\] (2)

2.2. Support Vector Machine

Support Vector Machine builds hyper plane or hyper plane set in high or unlimited dimension space, which can be used for classification, regression or other tasks. Intuitively, a good separation is
achieved by the hyper plane that has the greatest distance to the nearest training data point of each class (functional margin is called), because in general the larger the lower the margin the generalization error from the sorting. When the problem of origin may be expressed in a finite space dimension, it is often the case that in space, the set is not separated linearly. For this reason it is proposed that the finite dimensional space be mapped into a much higher dimensional space that might make the separation easier in that space.

The main idea of the SVM method is the concept of maximal hyper plane margin. With the discovery of maximum hyper plane margin then the vector will divide the data into the form of the most optimum classification. Some examples of hyper planes that may appear to classify data are shown by the following figure 1.

![Figure 1. Examples of multiple hyper planes](image)

From Figure 1 it is found that the H3 line (green) does not separate the two classes. The H1 line (blue) separates, with a small margin and a H2 line (red) with a maximum margin. Data classification is a common task in machine learning. Suppose that some given data points belong to one of the two classes, and the goal is to determine the class of a new data point to enter.

3. Design and Analysis

![Figure 2. Flowchart System Overview](image)

The overall system scenario will be explained as follows:
1. The dataset that has been collected with the file.txt extension will be entered into the database.
2. Datasets that enter into the database will enter into the preprocessing stage. The preprocessing stage is done with 5 processes is stop word removal, lemmatization, stemming, POS tagging and dependency parser. The result of the preprocessing stage will result in the preprocessing dataset.
3. The preprocessing dataset will be used at the feature extraction stage using the Double propagation approach. The result of this stage is the feature and opinion pairs data.
4. Determination of sentiment orientation will use feature and opinion pair data consisting of polarity determination and orientation of sentiment.
5. The result of determining the orientation of sentiment generates the opinion pairs and their orientation. Pair opinions and orientation can be used for the classification process.
6. The classification process is used to classify data testing into positive, negative or neutral oriented classifications.
7. The classification results are used for the generating process of the summary which will produce a summary of features and opinions along with their orientation sentiment.

3.1. Data Preprocessing

![Flowchart Data Processing](image)

**Figure 3. Flowchart Data Processing**

The preprocessing data scenarios performed (figure 3) consist of:
1. The preprocessing data to be performed will use the review dataset.
2. Lemmatization and stemming will use the Stanford Core NLP library.
3. POS tagging and dependency parser will use the Stanford POS Tagger and Parser library.
4. The result of preprocessing will produce a preprocessing dataset which will be used in the next step.

3.2. Product Feature Extraction

The feature extraction process using double propagation done can be seen in the flowchart as in Figure 4 below.

![Flowchart Double Propagation](image)

**Figure 4. Flowchart Double Propagation**

Scenarios of feature extraction using double propagation performed consist of:
1. The preprocessing dataset will execute feature based on rules in double propagation and as long as there are features found then the feature will be added to the tuples.
2. If there is no feature then the next product feature will be created for each review sentence.

3.3. Label Sentiment Orientation

The process of labeling the sentiment orientation can be seen in the flowchart as shown in Figure 5. Scenarios from the labeling of sentiment orientation are:

1. Label sentiment orientation is divided into 3 ie positive, negative, and neutral.
2. Labeling is based on the result of hand labeling (expert judgment). If on the dataset there is a symbol "+" it will be labeled positively, and "-" for negative.
3. Neutral label is given if not labeled by expert judgment and previously been preceded by a sentence review which already has label. If not respected by sentence reviews that have been labeled then will be skipped.

![Figure 5. Flowchart Labeling Sentiment Orientation](image)

3.4. Product Review Classification

The process of classification can be seen on the flowchart as in Figure 6 below

![Figure 6. Flowchart Process Classification](image)

Scenarios of the classification process are:

1. The classification process will use training data where the training data has been through the preprocessing stage, feature classification and sentiment orientation determination.
2. Training data will be conducted by using the training method of NBC and SVM.
3. Each training result of NBC and SVM will be tested using data testing that has been through the preprocessing stage and feature classification and performed simultaneously.
4. Classification results from NBC and SVM will result in a classification of NBC and SVM to be used in the next stage.

4. Experimental Result

4.1 Naïve Bayes Classifier Method

In this study used Naïve Bayes method which is a statistical classification that predicts the probability of a class. The Naïve Bayes algorithm can handle continuous learning, so this algorithm is appropriately used to handle data that often receives new input that has not been described in the previous training dataset. This method uses Bayes theorem approach, assuming each attribute is independent, has no relationship or dependence on each other, and calculates the probability of an event based on a particular condition.

Through the results obtained after experiments and implementation, it was concluded that the use of Naïve Bayes Classifier was considered good because the accuracy reached 79% and F-measure of 81% which can be interpreted according to the level of accuracy in Section 2.5.8 where 79% is in the range 0.80 - 0.90 and can be regarded as good classification and size of F-measure of 0.79 which approaches the score 1 which can be concluded that the performance evaluation of the NBC method in predicting or classifying data testing correctly can be classified well. And also the time execution reached 1363.8838 seconds with the composition of training data as much as 1800 and data testing as much as 429.

4.2 Support Vector Machine Method

The use of the Support Vector Machine method tends to produce better performance than the Naïve Bayes Classifier in terms of classification accuracy. Support Vector Machine is one method of classification with supervised learning. The SVM concept is described as an attempt to find the best hyper plane that serves as a class separator from several alternative hyper planes. This algorithm can be used for classification and regression as well as other tasks. Intuitively, a good separation is obtained by hyper plane which has the greatest distance to the nearest point on the training data of each class, because in general the larger the margin is formed, the smaller generalization error in the classifier.

Through the results obtained after experiments and implementation, it was concluded that the use of Support Vector Machine is considered good because the accuracy obtained has reached 85% and F-measure of 87% which can be interpreted in accordance with the level of accuracy in Section 2.5.8 where 85% is in the range 0.80 - 0.90 and can be regarded as good classification and the size of F-measure of 0.87 which approaches the score 1 which can be concluded that the performance evaluation of the SVM method in predicting or classifying the data testing correctly can be classified as well. And also the time execution reached 19.2162 seconds with the composition of training data as much as 1800 and data testing as much as 429.

Comparison of accuracy and time execution of both methods can be seen in the following graph.
5. Conclusion
The application of sentiment analysis can be used to classify product reviews into positive, negative and neutral categories. The simulator built by the researcher can handle every step done from the load dataset to produce a summary of the classification results.

Feature extraction is done using double propagation is in one review sentence can extract more than one feature of the product. The amount of data, composition and techniques of data selection training and testing can affect the accuracy and time execution. The methods used to classify product reviews are considered good and include good classification. Based on experiments that have been done, the results obtained that from the two methods used are Naive Bayes and Support Vector Machine in the classification, the accuracy of SVM method is higher than the NBC method and time execution SVM is shorter than NBC.

References

[1] A, G., R, B. & Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Volume I, p. 12.
[2] A, M., D, F. & F, P., 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. ICML 2000.
[3] Anon., 2010. Rotten Tomatoes Movie Reviews. [Online] Available at: https://www.rottentomatoes.com [Accessed 3 October 2016].
[4] A, R., 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. Proceedings of the Conference on Empirical Methods in Natural Language Processing.
B, P., L, L. & S, V., 2002. Thumbs up? Sentiment Classification using Machine Learning. *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia.*

Frank, E. & Witten, I. H., 2005. *DATA MINING Practical Machine Learning Tools and Techniques.* 2nd ed ed. s.l.:San Fransisco: Morgan Kaufmann Publisher,Elsevier Inc.

Garcia, D. & chweitzer, F., 2011. Emotion in Product Reviews – Empiris and models. Switzerland.

G, Q., B, L., J, B. & C, C., 2009. Expanding Domain Sentiment Lexicon through Double Propagation. *Proceedings of IJCAI.*

Han, J., Kamber, M. & Pei, J., 2012. *DATA MINING Concepts and Techniques.* 3rd ed. Elsevier : Waltham: Morgan Kaufmann..

Hu, M. & Liu, B., 2004. *Mining and Summarizing Customer Reviews.* Chicago: Department of Computer Science University of Illinois at Chicago.

J, M., n.d. *Amazon Review Data.* [Online] Available at: http://jmcauley.ucsd.edu/data/amazon/ [Accessed 2 October 2016].

Jurafsky & Dan, 2015. *Stanford | Natural Language Processing - Course Introduction.* [Online] Available at: https://class.coursera.org/nlp/lecture/124.

K, D., S, L. & D, M. P., n.d. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *Proceeding WWW '03 Proceedings of the 12th international conference on World Wide.* pp. 519-528.

Lackermair, G., Kailer, D. & Kannmaz, K., 2013. Importance of Online Product Reviews from a Consumer’s Perspective. *Advances in Economics and Business.*

Liddy, E. D., 2001. *Natural Language Processing.* 2nd Ed. NY. Marcel Decker, Inc ed. s.l.:In Encyclopedia of Library and Information Science.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. *Morgan & Claypool Publishers,* p. May.

M, H. J. & Kember, n.d. *Data Mining Concepts and Techniques.* Second ed. s.l.:s.n.

Ohana, B. & Tierney.B, n.d. Sentiment Classification of Review Using SentiWordNet. *Dublin Institute of Technology.*

Silwattananusarn, S. & KulthidaTuamsuk, n.d. Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012. *Thailand: Khon Kaen University.*

S, L. T., W, H. l. & A, H. T., 2011. Is Naïve Bayes a Good Classifier for Document Classification?. *International Journal of Software Engineering and Its Applications,* pp. 37-46.

Stats, I. W., n.d. *Indonesia Internet usage, broadband and telecommunications reports.* [Online] Available at: http://www.internetworldstats.com/asia/id.htm [Accessed 1 October 2016].

Vega, V. B., 2001. Information Retrieval for the Indonesian Language. *Master’s thesis, National University of Singapore.*

Zhao, Y., Yang, S., Narayan, V. & Zhao, Y., 2013. Modeling Consumer Learning from Online Product Reviews. *Marketing Science*