Support Vector Machine Classifier for Sentiment Analysis of Feedback Marketplace with a Comparison Features at Aspect Level

Hario Laskito Ardi
Department of Information System
Diponegoro University
Semarang, Indonesia
harios1si@gmail.com

Eko Sediyono
Department of Information System
Diponegoro University
Semarang, Indonesia
ekosed1@yahoo.com

Retno Kusumaningrum
Department of Informatics
Diponegoro University
Semarang, Indonesia
retno@if.undip.ac.id

***Abstract* — Sentiment analysis is an interdisciplinary field between natural language processing, artificial intelligence and text mining. The main key of the sentiment analysis is the polarity that is meant by the sentiment is positive or negative (Chen, 2012). In this study using the method of classification support vector machine with the amount of data consumer reviews amounted to 648 data. The data obtained from consumer reviews from the marketplace with products sold is handpone. The results of this study get 3 aspects that indicate sentiment analysis on the marketplace aspects of service, delivery and products. The slang dictionary used for the normation process is 552 words slang. This study compares the characteristic analysis to obtain the best classification result, because classification accuracy is influenced by characteristic analysis process. The result of comparison value from characteristic analysis between n-gram and TF-IDF by using Support Vector Machine method found that Unigram has the highest accuracy value, with accuracy value 80,87%. The results of this study explain that in the case of analysis sentiment at the aspect level with the comparison of characteristics with the classification model of support vector machine found that the analysis model of unigram character and classification of support vector machine is the best model.**

Keywords - Sentiments Analysis, Features Extraction,N-Gram, TF-IDF, Support Vector Machine, Marketplace, Aspect.

I. INTRODUCTION

Sentiment analysis is a field of study that analyzes one's opinion, one's sentiments, one's evaluation, one's attitude and one's emotions into written language. The technique of sentiment analysis can support many decisions in many scenarios. This research uses two attribute class, that is positive and negative, because in internet the comments that appear can be positive and negative comments. Consumer interaction is considered a valuable source of information because people share and discuss their opinions about a particular topic freely. The classification method that is now widely developed and applied is Support Vector Machines (SVM). This method is rooted in the theory of stati5t6stical learning which results very promising to provide better results than other methods. Many researchers have reported that SVM is probably the most accurate method for text classification. It is also widely used in semtiminal analysis. Aspect level is also called level feature [1].

In addition to searching for language constructs, the more aspect levels look at the opinion side. It is based on that opinion consists of sentiment (positive or negative) and target (opinion). From this level it can be seen that the importance of the opinion target serves to make it easier to understand the problems in the sentiment analysis. For large data sets SVM requires enormous memory for the allocation of the kernel matrix used. SVM training methods that require large memory are chunking and decomposition [2].

In sentimental analysis, the commonly used feature is n-gram. In some literature, it can also be interpreted the emergence of a new meaning or word from a set of characters cut in a word. Text feature extraction is an important step in text classification. Feature extraction plays a role in determining which features will be used by which classification techniques and which features are ignored. The large number of features resulted in a large dimensionality word vector. The addition of relevant n-gram features can improve text classification performance. In addition to the extraction of n-gram features, feature weighting, which is weighting the feature according to its significance, is a step that can be explored to improve classification performance. Commonly used weighting, Term Frequency - Inverse Document Frequency (TF-IDF), only considers the frequency parameters of feature appearance in the document and the number of documents containing the feature. SVM also works well on high-dimensional data sets, even SVMs that use kernel techniques must map original data from their original dimensions to other relatively higher dimensions. The kernel problem makes it possible to define nonlinear decision limits,

which are linear in the high-dimensional feature space, without calculating the parameters of the optimal hyperplane in the possible dimensional feature space [3].

Problems that exist in N-grams, char-n-grams or skip-grams produce "big feature" numbers, but not all features are significant for the classification process. The large number of features resulted in the generation of the "big dimensionality word vector" vector dimension. This can compromise the computation process either during pre-process, as well as subsequent processes (indexing, weighting and feature selection). On the other hand, the addition of relevant n-gram features can improve text classification performance [4]. Alternative features other than N-gram are TF-IDF. Basically TF-IDF works by calculating the relative frequency of a word appearing in a document compared to the inverse proportion of the word that appears on the entire document. In addition to the extraction of n-gram features, feature weighting, which is weighting the features according to their significance, is a step that can be explored to improve classification performance [5].

Examples of sentiment analysis scenarios are the utilization of data from social media such as twitter, product and facebook reviews combined with data from the company itself eg data from sales or customer data that already exist in the relational database [6]. Thus can be obtained analysis to perform a marketing strategy that is telling. For example by analyzing the people in influential social media to market the product [7].

II. RELATED WORK

A. Aspect level of Sentiment Analysis

Analysis done at document level and sentence level is still limited to the polarity of a sentence or document. But the two studies have not been able to know about what (objects) that people like or dislike. Aspect level is also called level feature. In addition to searching for language constructs, the more aspect levels look at the opinion side. It is based on that opinion consists of sentiment (positive or negative) and target (opinion). From this level it can be seen that the importance of the opinion target serves to make it easier to understand the problems in the sentiment analysis. Research at the level of documents and sentences is a challenge that is now the focus of researchers. For entity level and aspect is a difficult part, because this level consists of several other sub-issues. It can be concluded that sentiment analysis is a rapidly expanding field of science that can be explored further to obtain results that can be utilized in various fields, namely consumer products, services, health services and financial services for social activities and political election [8].

B. Preprocessing

In this study, the data to be used is Indonesian text data taken from the database twitter. Then the data to be processed is unstructured, therefore it is necessary to pre-process stages before the Support Vector Mechine. Preprocessing stages consist of Tokenizing, stop word removal, stemming and using weighting on every word throughout the document using the TF.IDF (term frequency-inverse document frequency) scheme [9].

1. Tokenizing

The Tokenizing process is useful for breaking every sentence of all the knowledge documents into words (term) using tab delimiters and spaced characters.

2. Stopword Removal

The quality of data mining methods such as clustering is very influential on the noise removal process used in the clustering process. For example frequently used words such as "the", may not be useful for improving the quality of clustering. Thus, it is important to choose the feature effectively so that the words noise can be eliminated before clustering [10]. The simplest way to select words in document clustering is the use of document frequency to filter out irrelevant words. In other words, words that often appear in documents can be omitted because they are plain words like "a", "an", "the" and "of" are not quite as diverse in terms of clustering. Stopwords are words that often appear in a less useful document in the process of extracting text. In Indonesian-language studies, stopwords used for example are "yang", "like", "is", "is", "a" and others.

3. Stemming

The purpose of stemming is to reduce the form of inflection and sometimes the word origin is related to the word in basic form. Stemming process is useful to change a word into its basic word, for example the word 'get' to 'can'. Stemming will improve the classification of texts in certain languages, in Indonesian, stemmer has been widely developed [11].

4. Normalization

Normalization is the stage of identification slang words and exaggerated words and then replaced with the word in the dictionary KBBI (Big Indonesian Dictionary). The steps performed by the normalization algorithm conducted in this study as follows [12] :

a. Search for words that will be normalized in the dictionary. If found then it is assumed that the word is root word.
b. If not found excessive lettering starts for each letter in the word, check the first letter of the word, then recoding. Check the next letter if the letter is the same as the previous letter then delete the letter, if not save the letters do the same thing in the next letter.
c. Doing recoding.
d. If it has been checked for each letter check the previous process hasis word on the dictionary.

e. If found then the algorithm stops, If not found this algorithm returns the original word before the excessive lettering is done.
f. Next check the word on the dictionary.
g. If found do change the word to standard word. If not found then return the word in the root word.

C. Feature Extraction

1. N-gram

N-gram is the cutting of longer words. In some literature can also be interpreted the emergence of a new meaning or word from a set of characters cut in a word (Trenkle and Cavnar 1994). Typically it is a single word piece into a set of overlapping N-Grams. The addition of underscores at the beginning and end of the word are used to help determine the initial state of the word and the end of the word. So in the word "TEXT" can be com- bined into the following N-grams:

Bi-gram : _T, TE, EK, KS, S_
Tri-gram : _TE, TEK, EKS, KS_,S_ _
Quad-gram : _TEK, TEKS, EKS_, KS_ _, S_ _ _ (1)

Therefore, a word with length k, added to the bottom line, would have k + 1 bigram, k + 1 trigram, k + 1 quadgram, and so on.

N-gram-based matching has been successful in handling unclear inputs such as, in interpreting postal addresses, restoring texts, and natural language processing applications. The key to successful matching based on N-gram is because each word is composed into small parts, the error that appears only affects a small number of parts, leaving the other one intact. If we calculate the same N-grams in two words, we will get a measure of the similarity of those two words that are not affected by various textual errors [13].

2. TF-IDF

Basically TF-IDF works by calculating the relative frequency of a word appearing in a document compared to the inverse proportion of the word that appears on the entire document [14]. TF-IDF consists of two component values namely term-frequency and inverse document frequency. This calculation can be used to find out how relevant the word is to a particular document. The formula for calculating tf-idf can be seen in equation :

$$tf\ idf(d,w) = tf(d,w) * \log \frac{n}{df\ w} \quad (2)$$

Where :

tf (d, w) : the frequency of occurrence of term w on document d
n : total number of documents
dfw : number of documents containing term w

Thus, the tf-idf value of a word will be high when the word is large in some small number of documents. So the word is very good to serve as a differentiator. Conversely, if the value of tf-idf in a low word means the word is in many documents, so less suitable to be differentiated.

2.4 Support Vector Machine

SVM (Support Vector Machines) are a useful technique for data classification. Although SVM is considered easier to use than Neural Networks, users not familiar with it often get unsatisfactory results at first.. Here we outline a "cookbook" approach which usually gives reasonable results. Note that this guide is not for SVM researchers nor do we guarantee you will achieve the highest accuracy. Also, we do not intend to solve challenging or diffi- cult problems. Our purpose is to give SVM novices a recipe for rapidly obtaining acceptable results. Although users do not need to understand the underlying theory behind SVM, we briefly introduce the basics necessary for explaining our procedure [15].

A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. Given a training set of instance-label pairs (xi , yi), i = 1, . . . , l where xi ∈ Rn and y ∈ {1, −1} l , the support vector machines (SVM) require the solution of the following optimization problem [16] :

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \quad (3)$$

Table 1 : Problem characteristics and performance comparisons.

Applications	#training data	#testing data	#features	#classes	Accuracy by users	Accuracy by our procedure
Astroparticle[1]	3,089	4,000	4	2	75.2%	96.9%
Bioinformatics[2]	391	0[4]	20	3	36%	85.2%
Vehicle[3]	1,243	41	21	2	4.88%	87.8%

Here training vectors xi are mapped into a higher (maybe infinite) dimensional space by the function φ. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameter of the error term. Furthermore, K(xi , xj) ≡ φ (xi) T φ(xj) is called the kernel function. Though new kernels are being proposed by researchers, beginners may find in SVM books the following four basic kernels :

- linear : K(xi , xj) = x T i xj .
- polynomial : K(xi , xj) = (γxi T xj + r) d , γ > 0.
- radial basis function (RBF) : K(xi , xj) = exp (− γkxi − xjk 2), γ > 0.
- sigmoid : K(xi , xj) = tanh(γxi T xj + r). Here, γ, r, and d are kernel parameters.

III. EXPERIMENT AND RESULTS

3.1 Experimental Data Collection

Consumer feedback data is taken from consumer marketplace reviews. The amount of data obtained amounted to 648 data feedback. The data is used to compile a slang dictionary and determine the dictionary aspect used for categorization. Once analyzed by linguists then obtained various aspects used for categorizing aspects of the marketplace. The analysis results obtained that in the marketplace there are 3 important aspects of delivery, products and services. Dictionary said aspect of delivery there are 100 words, dictionary word service aspect there are 117 words and dictionary of the word product there are 103 words. The slang dictionary used there is a 552 word slang that appears in 648 feedback made into a slang dictionary. In the making of this research dictionary involving linguists to conduct data collapse. After the data obtained, categorizing aspects and dictionary slang next step is preprosessing process. In the preprosessing process there are several stages of normalization, stopword removal and steamming. In the stopword removal process using literary and perspectives library based on data collected based on marketplace consumer reviews. The next step is stemming using algorithms from Nazief and Adriani that have been developed by literature. In the process of identifying word slang words that are not contained in the word dictionary will not be changed into the default word. The normalization process of replacing the word slang into a standard word that can simplify the process of stopword removal.

3.2 Experimental Procedure

Implementation of research procedures conducted in order to achieve the objectives to be achieved by performing several stages of preparation, design, programming, testing and improvement for the conclusion of the results of research undertaken. The first stage is preparation, this activity is carried out before the research is conducted by collecting literature such as international and national journals that are appropriate and supportive with research topics, as well as articles relevant to the research topic. Observations were made to cover the present problems, the current methods and the performance of the algorithm to be applied as input materials in the design of the information system to be built.

The second stage is the design stage that is based on the preparation stage by doing the workflow design. The third stage is the implementation of programming is preprocessing stages consisting of Tokenizing, Stopword removal, Lemmatization and Normalization. In the next stage is the characteristic analysis using 2 comparison of characteristic analysis by using TF-IDF and N-Gram. Stages of N-Gram characteristic analysis have 4 compositions: Uni-gram, Bi-gram, Tri-gram and Quad-gram. In the next process, it is the process of searching the classification model using SVM (Support Vector Machine) method. From the process of training data this time will get a new SVM model for test data. After that will get the result of positive and negative sentiment analysis.

The fourth stage is the improvement of the application of information systems built. Improvements are made to deficiencies found both at the input stage and at the process stage and output display. To get the conclusion of performance of wake up application calculation of errors (errors) from the results of information systems built that is by doing a comparison between the actual data value with the predicted value of the prediction of the period of time that has been set by using the equation that has been standard.

Development of information system begins with the determination of input (input) to be used, then proceed with the process as an implementation of the applied method and ends with the form of output (output) to be displayed.

3.3 Results and Analysis

The experimental result using K-Fold method with K = 10 got accuracy on each feature. Each feature is tested 10 times with total training and testing data of 648 data. After the calculation was obtained 65 training data with 9 times of test and 63 training data with one test.

Based on the result of research by using Support Vector Machine method in classification process and using 2 characteristic analysis to do comparison. Such feature analysis uses TF-IDF and N-gram. Such feature analysis will be compared to produce the best Support Vector Machine model of the sentiment analysis case based on consumer reviews on the marketplace. The following table describes the results of comparison research in table.

Table 2 : Table describes the results of comparison research

No	ACCURACY				
	TF-IDF	UNIGRAM	BIGRAM	TRIGRAM	QUADGRAM
1.	44, 6154 %	80 %	78, 4615 %	81, 5385 %	66, 15 %
2.	43, 0769 %	83, 0769 %	75, 3846 %	83, 0769 %	44, 6154 %
3.	43, 0769 %	84, 6154 %	78, 4615 %	60 %	61, 5385 %
4.	47, 6923 %	67, 6923 %	76, 9231 %	55, 3846 %	61, 5385 %
5.	47, 6923 %	81, 5385 %	73, 8462 %	53, 8462 %	56, 9231 %
6.	47, 6923 %	73, 8462 %	75, 3846 %	66, 1538 %	63, 0769 %
7.	43, 0769 %	92, 3077 %	69, 2308 %	76, 9231 %	58, 4615 %
8.	49, 2308%	83, 0769 %	67, 6923 %	78, 4615 %	55, 3846 %
9.	46, 1538 %	78, 4615 %	83, 0769 %	87, 6923 %	66, 1538 %
10.	49, 2063 %	93, 6508 %	76, 1905 %	66,6667 %	55, 5556 %
AVERAGE	46,416 %	80, 109 %	75, 594 %	72, 20025 %	55, 185 %

From the results of the above research we found that the accuracy of ngram is better than the accuracy of TF-IDF. Based on data transformation experiments from product review text of Ngram method to numeric integer value while tfidf method becomes numberi value of decimal number so that it influence to calculation process in SVM. Tabel 2 shows the accuracy value of each characteristic analysis in this study. So it is found that the Unigram - SVM model is the best model for consumer reviews case in the marketplace. Based on the result of research by using Support Vector Machine method in classification process and using 2 characteristic analysis to do comparison. Such feature analysis uses TF-IDF and N-gram. Such feature analysis will be compared to produce the best Support Vector Machine model of the sentiment analysis case based on consumer reviews on the marketplace. From the results of the above research we found that the accuracy of ngram is better than the accuracy of TF-IDF. Based on data transformation experiments from product review text of Ngram method to numeric integer value while tfidf method becomes numberi value of decimal number so that it influence

to calculation process in SVM. Tabel 2 shows the accuracy value of each characteristic analysis in this study. So it is found that the Unigram - SVM model is the best model for consumer reviews case in the marketplace.

IV. CONCLUSION

The result of comparison value from extraction method of characteristic between n-gram and TF-IDF by using Support Vector Machine method found that Unigram has the highest accuracy value, with accuracy value 80,87%. The results of this study get 3 aspects that indicate sentiment analysis on the marketplace aspects of service, delivery and products. The slang dictionary used for the normalization process is 552 words slang. This study compares the characteristic analysis to obtain the best classification result, because classification accuracy is influenced by characteristic analysis process.

References

[1] I. H. Witten, E. Frank, and M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). 2011.

[2] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining," *Forum Am. Bar Assoc.*, pp. 19–62, 2005.

[3] Ahlgen, P. Colliander, C., 2009. Document-document similarity approaches and science mapping : *Experimental comparison of five approaches. Journal of Informetrics 3. 49-64.*

[4] Al-Rowaily, K., Abulaish, M., Al-Hasan Haldar, N., & Al-Rubaian, M. (2015). BiSAL - A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security.

[5] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012

[6] Burges, J.C., 1998, A Tutorial on Support Vector Machines for Pattern Recognition, Kluwer Academic Publisher, Boston.

[7] Cao, Q., Thompson, M. A., & Yu, Y. (2013). Sentiment analysis in decision sciences research: An illustration to IT governance. *Decision Support Systems*, *54*(2), 1010–1015.

[8] Chang, C. and Lin, C., 2013, LIBSVM : A Library for Support Vector Machines, ACM Transactions on Intelligent Systems and Technology, National Taiwan University, Taipei.

[9] Cristianini, N. and Taylor, J. S., 2000, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press.

[10] Erra, U., Senatore, S., Minnella, F., & Caggianese, G. (2015). Approximate TF-IDF based on topic extraction from massive message stream using the GPU. *Information Sciences*, *292*, 143–161.

[11] Gaspar, R., Pedro, C., Panagiotopoulos, P., & Seibt, B. (2016). Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, *56*, 179–191.

[12] H. Li, Y. Chen, H. Ji, S. Muresan, and D. Zheng, "Combining social cognitive theories with linguistic ciri for multi-genre sentiment analysis," *Proc. Pacific Asia Conf. Lang. Inf. Comput.*, no. 1, 2012.

[13] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 54, no. Second Edition. 2006.

[14] Jurado, F., & Rodriguez, P. (2015). Sentiment Analysis in monitoring software development processes: An exploratory case study on GitHub's project issues. *Journal of Systems and Software*, *104*, 82–89.

[15] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," *Inf. Commun. Embed. Syst. (ICICES), 2013 Int. Conf.*, pp. 271–276, 2013.

[16] Pang, B. dan Lee, L., 2008, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, vol. Volume 2, no. Issue 1-2, pp. 1-135.

[17] Vapnik, V.N., 1999, The Nature of Statistical Learning Theory, 2nd edition, New York.