Feature-Based Sentiment Analysis in Online Review with Semi-Supervised Support Vector Machines (S^3VMs)

Jessie Setiady, Warih Maharani, Rita Rismala
Department of Informatics Engineering, Faculty of Informatics, Telkom University

Keywords:

Review
Sentiment
Product feature
S^3VMs

ABSTRACT

Online reviews provide facility so that internet user can give review about an aspect. Sentiments about a product are useful and have an influence in decision-making by person or organization. As in an opinion, reviewers and provide positive and negative reviews simultaneously. This is due, opinions targets are often not the product as a whole, but rather part of a product called the feature, where there are advantages and disadvantages in the eyes of reviewers.

In this research, sentiment will be identified based on its opinion. Opinion data used in this research is in English, taken from the site www.cnet.com. The product conclusions presented based on product features. Thus, there are two processes undertaken in this research: (1) Extraction of product features in opinion, (2) Sentiment identification for each product feature. Feature extraction is done by searching for phrases that match the relation dependencies template, and then do the filtering feature. In sentiment identification, the positive and negative probability value, and also the target class of the feature opinion, became S^3VMs input parameters. In the study by S^3VMs, some data are treated as unlabeled data. Results obtained from this study for the evaluation of sentiment identification with F1-Measure at 86% for positive class and 70% for negative class. As for feature identification obtained 82% accuracy. For further development of this research, Improve SVM is suggested to handle the unbalance data problem. Mapping to implicit feature is also advisable to identify more product feature.

Corresponding Author:

Jessie Setiady,
Department of Informatics Engineering, Faculty of Informatics,
Telkom University,
Jalan Telekomunikasi No.1 Ters. Buah Batu, Bandung, Indonesia.
Email: setiady.jessie@live.com

1. INTRODUCTION

Currently the Internet is not only used as a media to access information, but also as a media to share information. Information is categorized into two [5]: knowledge (facts), or opinion. Both types of information can be easily shared by Internet users, or which is known as User Generated Content, through a variety of facilities, such as: blogs, product review sites, social networking, forums, Question and Answer sites, voting sites, etc.

In the survey conducted by comScore (2007), and Horrigan (2008), found that 81% of internet users in the U.S. use the internet to search for products to buy, and more than 30% of internet users provide a review of a product purchased [1]. An online review is one medium that provides facilities so that a reviewer can give reviews or opinion, in the form of thoughts, suggestions or just comments. Reader, with their intuitive abilities, is able to know the sentiments of the reviewer of a topic of discussion, by browse the websites of online reviews available. Overview sentiment on reviewer opinion given can be used as one of the parameters of the analysis, such as the experience of others who have purchased a product determines a person's decision to purchase a particular product. The problem is, so many opinions are available, so that the reader will be overwhelmed if they have to read and analyze one-by-one reviewer's opinion. Another problem is, reviewers often provide an opinion on the positive and negative aspects of the product. Thus, an

opinion would be better if not simply be generalized into an expression of sentiment, but need to be separated by aspect or feature.

Based on these problems, it would be a benefit to the review searcher when there is an overview of the results of sentiment analysis of a product opinion. To achieve this, the opinion of the reviewer needs to be analyzed, identified and extracted features, and classified into class sentiment expressions such as: negative, or positive.

Currently, the method focuses on the paradigm of unsupervised learning, where all the data are not labeled, and supervised where all the data is labeled (for the training and evaluation). SVM included in the category of supervised learning, so in this case the overall opinion that the data will need to be labeled as a guide to determine the optimal position of separating hyperplane. However, the application of sentiment mining implementations, with many and varied opinions of data exist, it needs great effort and cost for labeling each of the data used for learning. Semi-supervised methods Support Vector Machines (S3VMs) used in this research. The expectation is, S^3VMs method can classify opinion into its sentiment expressions using combination of labeled and unlabeled data. The advantage is that the classification can be done though labeled data is scarce.

In this paper, presented the design and analysis for sentiment analysis based on its feature. The first part of this paper, presented the background and purpose of the study. The second part of this paper presented the overview of the previous study about these topics. The third part of this paper, presented the system design from opinion text as input data, transformed into information based on sentiment of its feature. In the fourth part of the paper presented results of the identification process with a typed-feature dependencies as well as the identification of its sentiment by using S^3VMs.

2. PREVIOUS STUDY

In the previous study [17], the method used to classify opinions into the sentiment expressions : unsupervised learning method using Pointwise Mutual Information (PMI) [23], a dictionary-based or lexicon-based [22], and supervised learning with machine learning methods such as Naive-Bayes Classifier, Maximum Entropy, and Support Vector Machine (SVM) [17]. In [17], compared with other machine learning methods, SVM method obtained the best performance.

As in [23], sentiment analysis is done using a semi-supervised method. In this study a semi-supervised methods that are used include: Self-trained Naive Bayes, Co-training, Expectation Maximization (EM) based SSL, and S3VMs. Through this study it was observed that the EM-NB consistently contributed well to the performance of the system, while S3VMs shows the reverse.

3. RESEARCH METHOD

The design of the system is generally illustrated in Figure 1. Based on the illustration, there are 4 major processes undertaken, include: Data Preparation, Feature Identification, Weighting, and Sentiment Identification.

The purpose of data preparation is to prepare the data to be processed by the classifier from data acquisition to stage the data labels. Steps being taken in the preparation of data include: Data Acquisition / Taking a review of reviews online content, and then perform the Symbol cleaning.

In Feature identification, opinion processed with Part-of-Speech tagger, so that the resulting opinion with word class tag. The Stanford parser used as a POS tagger. Once the dependencies words and word relationships known, we do filtering so that only the features that meet the relations contained in Table 1 are then determined as a feature. Explanation of the process is done at this stage include:

Tagging : using the Stanford tagger, every opinion annotated with the word class and word relation.

Extract frequent candidate feature : This stage is to extract features in opinion, which is adapted from [12]. Then do filtering using-typed template dependencies is adapted from [12], but there are a few additions, namely the handling of a negative, so the template used is as in Table 1 :

Filtering and grouping feature: do filtering feature using the threshold, so that only the features that often arise which are considered as features. Filtering is also performed on the features of synonyms, such as 'photo', 'picture', and 'image' are only considered as a feature of 'picture'.

Feature-based labeling: Any opinions which are separated by sentence labeled with class sentiment expression, positive (1) or negative (-1). Labeling imposed on the entire data manually by 3 people. Label the end of each opinion is the label of the

Table 1. Used Dependencies Relation Template

Template	Feature	Opinion
NN-amod-neg-JJ	NN	Neg + JJ
NN-amod-JJ	NN	JJ
NN-nsubj-neg-JJ	NN	Neg + JJ
NN-nsubj-neg—VB-dobj-NN	NN	JJ
NN-nsubj-JJ	NN	JJ
NN-nsubj-VB-dobj-NN	First NN	Last NN

most widely chosen by the giver label. S3VMs observation, some data will be treated as unlabelled data

VB-advmod-neg-RB	VB	Neg+RB
VB-advmod-RB	VB	RB

In Weighting stage, weights are determined by taking the probability of a positive and negative sentiment terms from the dictionary Sentiwordnet 3.0. Term used are trigram, bigram, and unigram. The weight of the calculation and the label will be on the classifier input features.

A sentiment features identified by S3VMs classifier. There are two stages, namely: Training, and Testing. In the training phase prior prepared training dataset with a dataset consisting of positive label (1), and dataset with negative labels (-1). Dataset then separated again, so that the resulting new dataset that contains the data to be treated as unlabeled data. Parameters accepted by classifier include: weights, labels (target), the kernel function and its parameters, and the C parameter which represents the upper limit (upper-bound).

The process is divided into 2 types:

Training process: Finding separating hyperplane / model using training data. Training data is a combination of labeled labeled data and unlabeled data are treated as the data (with the performance evaluation purposes)

The process of testing: Perform tests to testing the data. At this stage it will be measured by the level of its performance measurement parameters: precision, recall, and F1-measure, this parameter is used as an evaluation system.

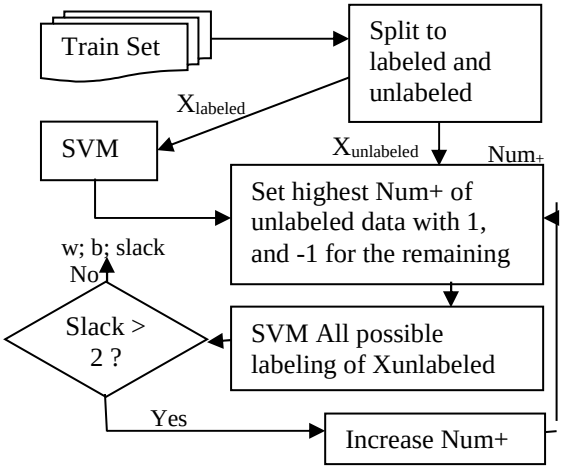

Figure 2. S^{3}VMs Process

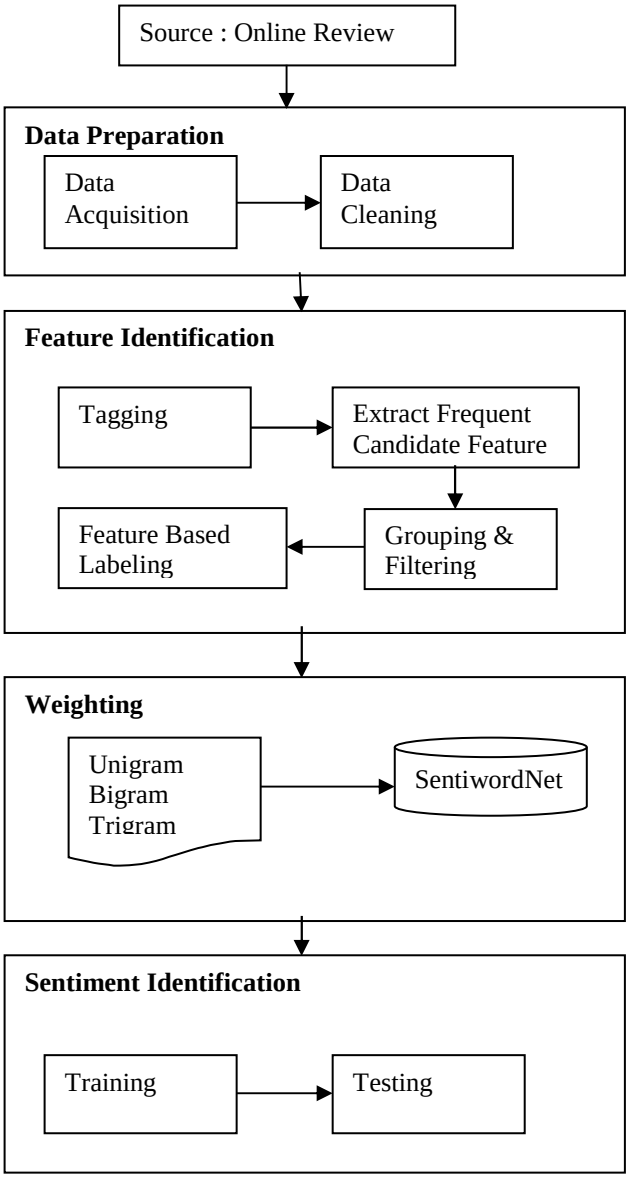

Figure 1. Research Global Process

4. RESULTS AND ANALYSIS

4.1. Feature Identification

Feature identification began by extracting features that are often mentioned in the opinion. Candidate feature extraction is done using templates that defined before. The feature candidates already pass the grouping stage. As for this research, there are 82 features that defined.

Table 2. Feature Identification Result With Threshold 1-10

No	Threshold	Feature Candidate	Non-Feature	Undetected Feature	Accuration
1	1	1405	1325	2	97.5%
2	2	362	282	2	97.5%
3	3	201	123	4	95.1%
4	4	128	56	10	87.6%
5	5	96	29	15	81.7%
6	6	73	6	15	81.7%
7	7	69	4	17	79.2%
8	8	52	1	20	75.6%
9	9	47	1	20	75.6%
10	10	44	0	21	74.3%

The percentage of accuration obtained from a reduced number of candidate features many non-features that are detected as a feature.

4.2. Sentiment Identification

To determine the effect of amount of labeled sample to the classifier S3VMs, selected data distribution 70:30, the constant c = 1, as well as the RBF kernel and gamma = 0.5. This is the combination of parameters that produces the best accuracy among the four test datasets SVM. The testing result shows in Figure 3.

The graph shows, at every decline in the number of data samples are labeled, also followed by a decrease in the value of F1-measure evaluation. However, the graphs also indicated that the data are not labeled are able to help so S^{3}VMs can find the right hyperplane. It is seen from the decline of the evaluation for POS classes based on accuracy and F1-measure, the overall data (100%) given the label, with the data that 40% of them do not have a target label, the difference in the largest decrease in F1-measure is only 0.88%, ie when the number of labeled samples is lowered by 10%. Whereas when the number of labeled samples derived respectively 20%, 30% and 40%, the percentage of F1-Measure for POS classes can reach values higher than 100% labeled data.

As for the NEG class data, the difference in the largest decrease when the number of labeled samples is reduced by 30%, the decrease in F1-measure of 1.392%. F1-measure the percentage had increased when the number of labeled samples is lowered by 20%, ie an increase of 0532%. However, the labeled sample decreased by 10%, 30%, and 40% indicated S^{3}VMs performance F1-measure is decreasing.

Figure 4 shows the processing time effect when the number of labeled sample decreased. As shown on the graph, processing time S^{3}VMs be higher when performing data processing labeled 90% and 80% labeled data. But it was able to process faster than SVM processing time when process data labeled 70% and 60% of data labeled. However, when seen from the processing time, the maximum difference between the processing of the SVM and S^{3}VMs of only 1.07890708 seconds. Based on these results it can be concluded that the classification by using S^{3}VMs only slightly sacrificing processing speed.

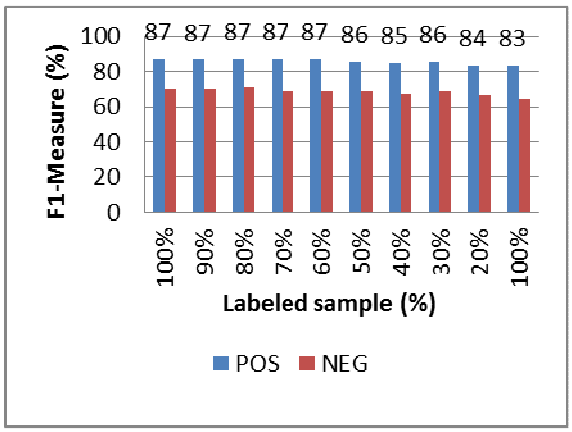

Figure 3. The result for amount of labeled data scenario, compared to performance in F1-measure percentage

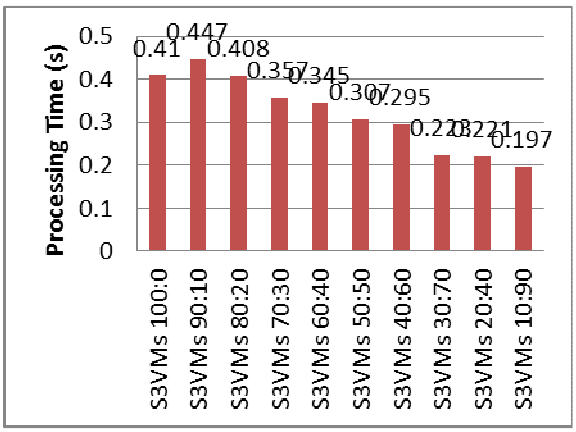

Figure 4. The result for amount of labeled data scenario, compared to processing time while training

5. CONCLUSION

Based on the research that has been done, resulting conclusion as follows: Determination of the sentiment expression of feature-based opinion using all labeled data resulting F1-measure reaches the highest obtained with the selected class is 86% positive. As for the negative obtained the highest F1-measure is 70%. This shows the SVM good to classify positive sentiment, but the performances decreased when classify sentiment classes. This could be due to the characteristics of the unbalanced dataset with negative data only about one-third of the overall data. The parameters used SVM is a constant c = 1, kernel = RBF, and gamma = 0.5.

Balancing techniques with down-sampling of data is observed not help increase SVM evaluation results significantly. The classification results on balance the data obtained F1-measure testing for positive class by 75%, and negative class by 76%, which are not better results when compared with the classification of data unbalance.

Semi-supervised classification using SVM can produce the observed performance in terms of processing speed and performance in terms of accuracy, precision, recall and F1-measure are not much different. On classification using S^3VMs with data reduction by 40% labeled F1-measure results obtained for the class of 85 145% positive, F1-measure which is the highest margin of only 1.576354% of the F1-measure on SVM classification using the entire data labeled. The labeled data reduction of 40% means that the amount of unlabeled data is 406 of the total 1450 data.

Classification using a Semi-supervised SVM observed an increase of the time, but not very significant. Reasonable time increase occurs because the increased computing is done when the semi-supervised SVM to process unlabeled data. In the study in this reserach, the observed difference between the maximum increases in processing time is equal to SVM and S^3VMs 1.07890708 seconds.

With such results, it can be concluded that a feature based sentiment classification using S3VMs can produce equally good results when compared to classification using SVM. Classification by using S3VM also more favorable in terms of time (to provide data labels from expert user), and can directly impact on cost savings. However, classification using S3VM also has a weakness, which can result in long processing time if the data are not labeled very large (too much repetition in S3VMs process), it is difficult to be optimized [24], and can get stuck in a local optimum (due to the step : check slack> 2).

ACKNOWLEDGEMENTS

Authors wish to thank Mrs. Warih Maharani and Mrs. Rita Rismala for guidance and feedback during this research work. Authors also thank to Stanford Parser researcher at Stanford University for provide a very useful program for this research.

REFERENCES

[1] L. Barbosa, *et al*., "For a few dollars less: Identifying review pages sans human labels", *NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 494-502 , 2009.

[2] KP Bennett and A. Demiriz, "Semi-supervised Support Vector Machines", *Proceedings of Neural Information Processing Systems*. 1998.

[3] Z. Ceska Z and C. Fox, "The Influence of Text Pre-processing on Plagiarism Detection", *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1221-1230, 2009.

[4] P. Chaovalid P and L. Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", *In: Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[5] L. Dey and HSM. "Opinion Mining from Noisy Text Data", *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 205-226, 2009.

[6] R. Feldman and J. Sanger, "The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data". *New York: Cambridge University Press*, 2007.

[7] D. Hand, H. Mannila and P. Smyth, "Principles of Data Mining", *The MIT Press*, 2001.

[8] JA. Holyst, *et al*. "CyberEmotions: Collective Emotions in Cyber-Space"., *The European Future Technologies Conference and Exhibition*, 2011.

[9] F. Jing, L. Zhuang and Zhu Xiao-Yan. "Movie Review Mining and Summarization". *Microsoft Research Asia, Department of Computer Science and Technology, Tsinghua University Beijing, P. R China,*

[10] T. Joachim, "Making large-scale SVM Learning Practical", *Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed)*, 1999.

[11] I. King and Z. Zu, "Semi-supervised Support Vector Machine", *In: Basics of Semi-supervised Learning. Hong Kong: ICONIP*, 2011.

[12] B. Liu, "Sentiment Analysis and Subjectivity", *In: Liu B. Handbook of Natural Language Processing. 2nd ed. Chicago*, 2010.

[13] PM. Marcus. B. Santorini and MA. Marcinkiewizz, "Building a Large Annotated Corpus of English : The Penn Treebank", *Compational Linguistik*, pp. 313-330, June 1993.
[14] MY. Nur and DD. Santika, "Analisis Sentimen Pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine". *In: Konferensi Nasional Sistem dan Informatika*, Bali, 2011.
[15] B. Ohana and B. Tierney, "Sentiment Classification of Reviews Using SentiWordNet", *Computer Science Common*, 2009.
[16] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques". *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol 10, pp. 79-86, 2002
[17] Y. Permadi, "Kategorisasi Teks Menggunakan N-Gram untuk Dokumen Berbahasa Indonesia". Institut Pertanian Bogor. 2008.
[18] AM. Popescu and O. Etzioni, "Extracting Product Features and Opinion from Reviews", *Proceesing of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, 2005.
[19] C. Soumen, "Mining The Web : Discovering Knowledge From Hypertext Data", San Francisco: Morgan Kaufmann Publisher, 2003.
[20] B. Santosa, "Support Vector Machines". *In: Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis*.,2007
[21] M. Taboada, J. Brooke, M. Tofiloski and KSM. Voll, "Lexicon-based Methods For Sentiment Analysis", *Computational Linguistic*, 2011.
[22] PD. Turney, "Thumbs up or thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews". Vol III. 2011.
[23] Y. Ning and K. Sandra. "Semi-supervised Learning for Opinion Detection", *Indiana University*, 2010.
[24] Zhu and Xiaojin. "Semi-supervised Learning Tutorial", *In : Department of Computer Sciences. University of Wisconsin*, USA, 2007.

BIBLIOGRAPHY OF AUTHORS

Jessie Setiady was born in Bandung, Indonesia, November 30th 1989. She have been working in Information Technology for 2 years. She graduated from Telkom Institute of Technology, majoring in Informatics Engineering (diploma), and then continuing studies in Bachelor Degree Informatics Engineering- Telkom Institute of Technology. Her interest in research on text mining and information in social media based on experience utilizing the facilities and information on social media

Warih Maharani was born in Semarang, Indonesia, March 24th 1978. She spent her youth in Semarang until 1994, and then she moved to Bandung and still. She graduated from Bachelor Degree Sekolah Tinggi Teknologi Telkom (now Institut Teknologi Telkom), majoring Informatics Engineering. She also graduated from Master Degree Institut Teknologi Telkom majoring Telekomunication Enginering. She become Lecturer in Institut Teknologi Telkom since 2002 to present. Her research interest is Information Retrieval and Text & Social Media Mining.

Rita Rismala was born in Ciamis, West Java, Indonesia, December 15th 1986. She moved from Ciamis to Bandung at 2005. She graduated from Bachelor Degree Insitut Teknologi Telkom at 2010 and then she began work in IT company for a year. She then continued her study to master degree Institut Teknologi Telkom at 2011 majoring in Data Mining – Informatics Engineering. She also become a Lecturer in Institut Teknologi Telkom from 2011 to present.