Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique

Conference Paper · May 2016

DOI: 10.1109/ICoICT.2016.7571912

CITATIONS
14

READS
853

3 authors, including:

Cut Fiarni
Institut Teknologi Harapan Bangsa
18 PUBLICATIONS **47** CITATIONS

SEE PROFILE

Sentiment Analysis System for Indonesia Online Retail Shop Review Using Hierarchy Naive Bayes Technique

Cut Fiarni[1], Herastia Maharani[1] ,Rino Pratama[1]
cutfiarni@ithb.ac.id, herastia@ithb.ac.id, rinopratama@ithb.ac.id
Department of Information System [1]
Institut Teknologi Harapan Bangsa (ITHB)
Bandung, Indonesia

***Abstract*—The rapid growth of internet user and the popularity of social media network has changed how people interact and doing everyday activities. Indonesia Small Medium Enterprise Organizations, such as the retail industry has also started to uses various social media to market their product online. The rapid growth of internet user and the popularity of social media network has led to big data of online opinion. Analysis on these opinions is very important because it can extract knowledge that can be the basis in making business decisions for the organizations. The problem is Indonesian citizen communicate in Bahasa and local languages, not to mention slang languages. So to build a sentiment analysis system is not easy because it has to be able to identify words and classify its sentiment. To overcome this problem, a sentiment analysis system that able to process opinions from social media using text mining is developed. This proposed approach would use feature extraction and selection to select words from learning dataset of Indonesian corpus and then classifying them to the respective class of target objects and sentiment. Then, we adopt the Naïve Bayes Classifier technique, with 3 sentiment classifications, aspects of online retail shop, and polarity of sentiment (positive, negative and neutral) and the polarity of the aspects of online retail shop. Results from this study shows that the sentiment analysis system for clothing product on social media using Naïve Bayes Classifier method is able to classify user opinions with 97.25% precision, 89.83% recall, and 89.21% accuracy.**

Keywords—Big data; Sentiment Analysis; Feature extraction, Indonesian Corpus, Naive Bayes Classifier.

I. INTRODUCTION

Today the rapid growth of internet user has changed how people interact globally, especially in doing business. The implementation of internet on business transactions has created new opportunities on how product or services sale in today's world. The emergence of online social networking (OSNs) has given internet users a medium for expressing and sharing their thoughts and opinions on all kinds of aspects of their life, including products and services. These phenomena also happen in Indonesia, as the fourth most populous country in the world, with internet penetration 22,1%, and internet user 55 million people and Facebook as the most popular website to access [1].

Online shopping as part of e-commerce refer to the process of selling products or services over the internet. In Indonesia, online shopping is becoming increasing popular because of its speed and easy to use. Two tops popular online retail shop are www.kaskus.co.id and www.tokobagus.com , in which both are Indonesian web sites and used Bahasa to interact [1]. OSNs, such as Facebook and twitter , with massive amount of comment and twit daily has become a gold mine for organizations to monitor their product's brand, reputation and also to understand their target market needs, by analyzing the opinion of their costumer's tweets or Facebook comment.

However, to evaluate those online feedbacks, is not a simple matter. Sometime when analyzing these fast growing online reviews, it becomes difficult to categorize whether costumer opinion are satisfied or dissatisfied of the products and service. Moreover, as part of improving their quality, organizations such as online shop need to classify what aspect of products and service that costumer most incline of. And because of the value of this information, especially to maintain costumer trust and retention, it all must be done in a timely matter. The problem is about 85% of text available on the internet has an unstructured format, so it needs to develop a system that can automatically classify aspects and sentiment from the online text data. [2]. Moreover, each online reviews not only discuss one subject matter, it could discuss various aspects with different range of sentiment (positive, negative or neutral). Sentiment analysis could solve this problem, by analyzing the emotion and context of the given online feedback. In this research, we will design an automatic sentiment analysis system for online retail business, specifically 'Distro', which are very popular among Indonesian youth.

ISBN: 978-1-4673-9879-4

Distro is a shortening of the Indonesian word "distribusi", or "distribution". Generally distro differ from other youth fashion outlets by their links to the independent music industry, hobbies, and other popular Indonesian young-adult lifestyles. Distro products and goods are very unique and made in small runs. Originality and scarcity becoming the selling points. Today, Distro also marketing and sales their products and good through the internet and Online social networks (OSNs) such as Facebook, twitter and so on. This research will focus on Indonesia's retail online shop (Distro) that sell clothing, accessories and shoes, that activities reach 70% aver all Indonesian e-commerce activities [1].

The purpose of this research is to study, analyze and implement the most suitable techniques for Sentiment Analysis System for Indonesia Online Retail Shop Review. In this proposed approach we present a supervised technique to classify online review based on the most important aspect on online distro shop and the sentiment behind those online reviews. The rest of the paper is organized as follows. In section 2 we will discuss related work and overview some techniques which are commonly used in Sentiment Analysis. Then, in section 3, we will discussed for the proposed sentiment analysis approach and technique. Section 4 describes the design and implementation of the proposed sentiment analysis system. Last, the section 5 covers evaluation and results of the proposed approach, and finally section 6 concludes the outcome of the research.

II. RELATED WORK

Sentiment analysis, also known as opinion mining is the methods used for enabling computers to recognize and classifying opinions from big unstructured texts datasets with machine language and computer programming. Its main purpose to determine the context and emotion of online text data. Sentiment analysis, concept and techniques first introduced by Liu, B. He defined sentiment has quintuplet aspect, which are: target object, a feature of the object, the sentiment value of the opinion of the opinion holder, the polarity of the opinion and opinion, opinion holder and the time when the opinion is expressed. [3]. The basic problem of opinion mining is opinion extraction. It is required to know the linguistic terms and get the idea from the text classification of subjective and objective terms identified by syntactic features. Another main focus is on subjectivity detection. Subjectivity is used to express the context or specific domain of online reviews [4].

Today, sentiment analysis has become a large and growing field. In this sentiment analysis approach we used supervised methods, because its nature to be generally more accurate than unsupervised approaches. But On the other hand, this method requires labeled training data and the goal is to classify an online reviews as referring to one or more of the aspects. The research to Classification sentiment has been done by another researcher with various approaches and algorithm. Most of the algorithms for sentiment analysis are based on a classifier trained using a collection of annotated text data. Before training, data is preprocessed so as to extract the main features. Some classification methods have been proposed: MaxEnt, NBC, SVM, KNN, etc. However, because its varied result in various research, it is still not clear which of these classification strategies are the more appropriate to perform sentiment analysis system [5].

In this study, we design a sentiment analysis system to help gain information and knowledge regarding Indonesian online costumer's opinion towards Distro's products and services using Naïve Bayes classifier (NBC). NBC is a probabilistic learning algorithm that derives from Bayesian decision theory.

NBC would combine previous knowledge with new knowledge. This classification algorithms are simple has performance similar to other approaches. In NBC, the probability of a message d being in class c, P(c|d), is computed as shown in these equation formula:

$$P(c|d) \propto P(c) \prod_{k=1}^{m} P(t_k|c), \quad (1)$$

Where P(tk|c) is the conditional probability of feature tk occurring in a message of class c and P(c) is the prior probability of a message occurring in class c [6].

III. RESEARCH APPROACH

In this section we will describe the approach for Sentiment Analysis System for Indonesia Online Retail Shop using Hierarchy Naive Bayes Technique.

A. Problem Formulation

In this section we will discuss the problem formulation for the proposed research approach. Indonesia is the world's fourth most populous nation, with over 250 million citizens and speaking in Bahasa as official language and 654 local languages identified [7]. In day by day communication, especially via internet, Indonesian citizen uses Bahasa and nearly 20 the most popular local languages, not to mention, of slang word and "alay" words as today most popular language among Indonesian youngsters. This condition resulting a challenging task to design and develop and automatic system. Moreover, to analyze sentiment of the comment need previous knowledge to classify the polarity of the opinion. For example, In Bahasa, the words 'Bagus', 'baik', '' keren' can be marked as having positive polarity, while the words 'jelek', 'parah' can be marked as having negative polarity.

In order to enhance the accuracy of sentiment analysis system, we propose a supervised model based on training corpus in order to build a dictionary

that contain related words to Indonesian Online Shop Aspects and their polarities.

Franky et al define subjectivity lexicon as the list that contains adjective words together with their positive and negative polarity information [8]. Subjectivity lexicon can be categorize as a general-purpose one or a domain-dependent or context-dependent one [9]. A domain-dependent or context dependent lexicon is a lexicon that is targeted to a specific domain or context. For example, in a sentence like 'suka sm warnany :) qra" wat cwe ada yg ky gini ??' (like the color, is there for women??), which the occurrences of words 'suka ' (like) consider positive, but we need to context or domain of the sentiment, in order to gain more knowledge of costumer need. In the example, human logic could easily classify the comment as positive sentiment for aspects of product, and there is a need of woman version of the said product. Moreover, from the same example, the words "cwe" is a slang words for women, so to enhance proposed system accuracy and precision, we would build a dictionary for slang words and replacement to correct the misspelled word and categorizing them with this synonym.

In this work, we experiment with subjectivity lexicons that are simple and only contains words and their polarity, and also to specific domain, which are Indonesian Distro. A domain specific lexicon focuses on collecting words that are generally acceptable by human as being positive, negative or neutral is needed, because in there are adjective words that related to specific aspect of distro, instead of combination of noun and adjective, like murah (cheap) specific to price aspect, or adem (cool) specific to material aspect of product and so on.

B. Dataset Collection

The first step before we could present a method for classifying the aspects and sentiment of the online retail business review are dataset collection and text preparation. Because our proposed sentiment analysis system adopts a supervised method, building an annotated corpus is really important for the learning process. In this proposed approach we used prior online reviews from OSNs of " XYZ Online Distro page" as knowledge to classify the aspects and sentiment of the online retail business.

XYZ Distro is one of the most popular distro among Indonesian youth today. XYZ distro, which already has several branches in various major cities in Indonesia, is also utilizing internet as a means of marketing and sales of its products. Because of the supervised learning of this approach, the success rate of this method depends on initial knowledge, in this case datasets from online review and feedback of XYZ distro on Facebook. This dataset contains more than 1442 online reviews in the time period from January 1st 2015 to June 23rd 2015. Then we divided the collection data into two groups of the dataset. Each group, dataset is for training the appropriate sentiment analysis approach for this research and for testing the proposed approach. After further analysis on the validity and content from each online review, we reduce the training dataset from 651 to 217 online review.

C. Proposed Method

In this research, we adopt of Liu's aspects of sentiment analysis and make some adjustment to accommodate Indonesian language. We will build a model for Indonesian online distro classification opinions in three major groups, which are:

- Target object
  This group would give knowledge regarding which aspect of online distro is most popular and what not.
- Polarity of the sentiment
  This group would give knowledge regarding polarity of the opinion. Usually in this group we use Indonesian adjective words and classifying them as negative, positive and neutral sentiment.
- Polarity of the target object
  This group would give knowledge on sentiment of each aspect on online distro.

Our system is divided into two subsystems as describe on Figure 1. The first subsystem is learning process, so the aim of the first subsystem is to build models for the proposed system. The second subsystem goal is to identifym and determine the polarity of the input dataset. This subsystem uses to test the proposed model that has been built on the first subsystem, by doing the classification process on input dataset.

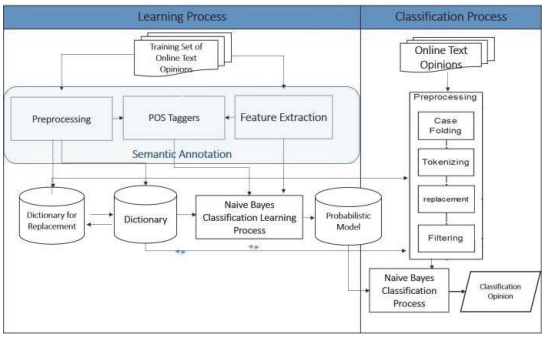

Fig. 1. Research Overview Diagram

In order to produce the right model that has a high accuracy level, we conduct learning process in three stages, which are:

1) Pre-processing

In this stage, the input data set to undergo several processes as described in figure 2. In this stage first, we are changing up all uppercase to lowercase in the input train dataset. Then splitting the sentence into words or token, this process called tokenization. Then we built dictionary to replace token that related to online distro aspects and sentiments. Since online review prone to misspell and grammatical error, this process becomes crucial. Moreover, as mention on the previous section,

Indonesian people use Bahasa (official language), local languages and slang terms in their everyday communication, the more big replacement dictionary is the more powerful the proposed model will be. Last, Filtering is the phase of removal the token that is not considered to contain any meaning or thought there should be related to online distro aspects and sentiment. This last phase would need knowledge that gathers from feature extraction. Table 1, illustrate these pre-processing stage. The result of pre-processing of classification learning process is matrix table as illustrate on table 2.

TABLE 1. ILLUSTRATION OF PRE-PROCESSING STEPS

Online review	Mantap nech XYZ...kemaren ane kesana pas lagi pasang lemari penitipan barang....jadi nyaman belanjanya Pelayanan manttap...dengan sabar dan sopan melayani pelanggan...apalagi ada tempat duduk untuk nunggu,,,wah makin nyaman dah....monggo cek TKP kalo gak percaya
Case folding	mantap nech xyz kemaren ane kesana pas lagi pasang lemari penitipan barang jadi nyaman belanjanya pelayanan manttap dengan sabarrr dan sopan melayani pelanggan .....
Tokenization	mantap-nech-ouval-kemaren-ane kesana-pas-lagi-pasang-lemari-penitipan-barang-jadi-nyaman-belanjanya-pelayanan-manttap-dengan-sabarrr-dan-sopan-melayani-…...
Replacement	Manttap = mantap, sabarr=sabar
Filtering	mantap, nyaman, mantap, sabar, sopan, melayani, nyaman

TABLE 2. MATRIX OF DISTRO'S ASPECT AND SENTIMENT

Input Dataset	**Aspect**				**Sentiment**		
	1st category		***n category***		+	-	***neutral***
1st review	Word 1	word 2.	.....	Word n	Word 4	Word 5	Word n
2nd review	2	1	.....	.....	2	1	.....
n	.....	n	n	.....	.....	n	n

2) Feature Extraction

Because of the nature of online review usually highlight more than one dimension of the distro's target object and has more than one polarity of sentiment, this lead to the need for dimensional reduction for distro's target object. Dimensionality reduction is one of the most crucial stages of built sentiment analysis model in order to remove irrelevant and unrelated comments. Both Feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better models, and built an efficient storage. In this research, we used feature extraction and selection to select words from learning dataset of online review and then classifying them to the respective class of target objects and sentiment. So dimensionality reduction is done by selecting features that are capable of discriminating words (token) that belong to different classes.

There are two major steps in this proposed approach. First is to build aspect and sentiment class, and the second is to build an annotated corpus based on their respective class. In the first step we need to identify an overall Distro character which represents one of six store dimensions: product assortment and variety, value of the merchandise given its price, service, location, facilities, and store atmosphere [10]. In this first step we built 8 class for online distro aspects, which are ; Bahan (material), Produk (product), Harga (price), Kualitas (quality), desain (design), Pelayanan (service), Ruang Pameran (Room Exhibition), Umum (general).

The second step is to build an annotated corpus based on their respective class. In this stage, first we analyze all important aspect related to online Distro in order to get targeted object's class. Afterward, we extract all of the aspect terms related to each class of online Distro classification from all token that we get from the preprocessing stage. This token usually noun and other predefined words. Then we calculated the number of tokens per opinion (Bag of Word) and count the number of keywords into a particular label. Table 3 shown a list of target objects and number of keywords that we get from the input data set. The same process goes to extract the polarity of sentiment (positive, negative and neutral) and the polarity of the target object. Then, in this stage, we map this class to 103 adjective token into respective class. For polarity of target object, we combine token into two subsets which are adjective, noun and predefined words.

3) Naive Bayes Classification Learning Process

This learning process goal is to get probabilistic value for each word on each classification domain group as mention above. In the proposed system, we adopt naïve Bayes to classify existing opinion. There are several steps this calculation process:

- Calculate probability for each class of Indonesia online retail shop aspects
- Calculate likelihood probability
- Calculate the highest probability of Distro's aspect and sentiment

In this stage we calculate $P(V_j)$ for each class category using the formula:

$$P(V_j) = \frac{|fd(V_j)|}{|D|} \qquad (2)$$

Where $fd(V_j)$is the number of words in the category j and D is the number of documents used in training. Furthermore, we calculate $P(W_k|V_j)$ for each W_k in the vocabulary with formula:

$$P(W_k|V_j) = \frac{f(W_k|V_j)+1}{N+|W|} \qquad (3)$$

Where $P(W_k|V_j)$ is the amount of occurrences of

word W_k in the category V_j, N is the amount of all words in the category V_j and |W| is the number of unique words (distinct) on all training data [6].

From learning process we get annotation corpus and matrix of Distro's aspects and sentiment, that we used to build proposed probability model that would be used on classification in three major groups as explained on previous section. From learning process we also build three database, first database of dictionary that contain vocabulary and adjective word and also a replacement dictionary. Replacement dictionary contain list of synonym and set of arrangement for Indonesian slang and alay words. As explained in the previous section Indonesian Citizen use Bahasa, various Local Languages and Slang words in their day by day communication. So in order to increase system precision and accuracy, in the proposed sentiment analysis system, we built incremental replacement dictionary. With this function, user could add list of new words respectively to their synonym in the replacement dictionary. Moreover, the proposed system also could visualize the sentiment analysis result. In the form of dashboards. This dashboard also has function to search and select data based on their time line and OSNs origin. This two function become important feature of the proposed system.

TABLE III. LIST OF DISTRO'S TARGET OBJECT

Target Object Label	Sum of token from input data set (n)
Bahan	9
Produk	42
Harga	3
Kualitas	2
Design	10
Pelayanan	7
Room Exhibition	7
Umum	42
Total	122

IV. SENTIMENT ANALYSIS SYSTEM OF INDONESIA ONLINE RETAIL SHOP

In this section we will describe the feature on the proposed system. In this development phase, we analyze the functional requirement, user interface design and implement the proposed system as illustrate on figure 1, in orderly phase. This process includes not only the actual writing programming code and train data set, but also the preparation of requirements, the system design, and confirmation that the proposed Indonesia Distro Sentiment Analysis System has met the research objectives. The basic functional requirement of the proposed system are:

- Authentication of user
- Import data from Facebook and twitter for training and classifying sentiment analysis purpose
- Range time of Import data
- Classify online opinion in three major group, which are : Target object, Polarity of the sentiment and Polarity of the target object
- Dashboard of Sentiment Analysis system, this feature would visualize Sentiment Analysis Results based on their target object, polarity of sentiment and polarity of target object.

The next step of system development is interface design of proposed system. Interface design is the process of defining how the system will interact with external entities pleasantly. Figure 2 shows the user interface for learning process for replacement dictionary. Identification of keyword is important for sentiment analysis because it could increase system ability to process and identify words.

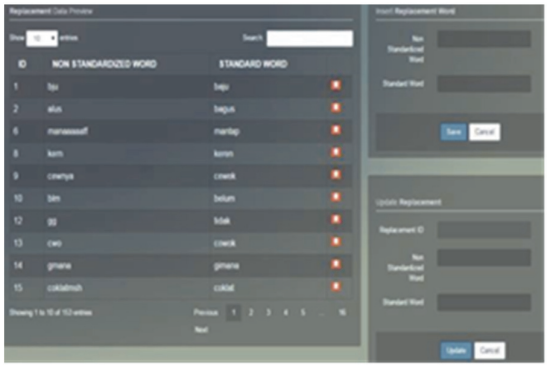

Fig. 2. Page of Replacement Keywords

Figure 3 shows the dashboard of sentiment analysis for specific time range. This feature also visualizes the sentiment according to three group as explained on previous section. As illustrate on figure 3, the proposed system also visualize the most discuss target object and the object and polarity of sentiment, but also a polarity in their relationship with target object. This feature also used to get the most discuss aspect and the overall polarity of online opinion of XYZ distro's costumers, and can be used as strategic marketing decision.

Fig. 3. Dashboard of the Proposed Sentiment Analysis System

To analyze the system capability to extract and classify opinion in to their respective class, we analyze the proposed system using testing dataset. The amounts of testing data that used are 217 comments. The results of the proposed system are:

$$Recall = \frac{ce}{ce+te} x\ 100\% \quad (4)$$

$$= \frac{195}{217} x\ 100\% = \mathbf{89.86}\%$$

$$Precision = \frac{ce}{ce+fe} x\ 100\% \quad (5)$$

$$= \frac{211}{217} x\ 100\% = \mathbf{97.24}\%$$

Where *ce* is the number of entities extracted correctly, *te* is the number of true entities not extracted, and *fe* is the number of false entities extracted.

This precision and recall numbers shown that the proposed system able to process and classify opinion in their respective class with high accuracy. The processing time of the classify process is 2.4 seconds, so this shown that the system can handle more data and processing it in short processing time.

V. CONCLUSION

In this proposed system, we used Naïve Bayes Classifier (NBC) in order to get sentiment and aspects classification of online retail business. With this approach we used prior online reviews from OSNs of Online retail business as a knowledge to classify the aspects and sentiment of the online retail business. This research utilizes NBC technique to gain knowledge from online opinion and categorizing them in three major groups, which are: Distro most discuss Aspects, Polarity of sentiment and Polarity of distro aspect from online opinion. In this proposed system, we built 8 class for online retail aspects, which are ; Bahan (material), Produk (product), Harga (price), Kualitas (quality), desain (design), Pelayanan (service), Ruang Pameran (Room Exhibition), Umum (general). The level of accuracy of the proposed Indonesian distro sentiment analysis system has precision accuracy rate of 97.24% and recall of 89.86% accuracy rate. This shown that the proposed sentiment analysis system is able to process online opinion with high data compatibility. This study will help organization, especially distro and other small and medium enterprise industry in Indonesia to have a better understanding of their target market needs and wants. The future work of this research is to build the incremental sentiment analysis system that could train new keyword in to existing classes and also to expand new group on online retail aspects in order to increase the system's ability to classify and analyze online opinion.

REFERENCES

[1] Ministry of Communication and Information, "2012 Indonesia ICT White Paper," 39-40; Kominfo, "Laporan Akhir Tahun 2013," December 27, 2013

[2] Reddy V, Siva RamaKrishna, et al, " Classification of movie reviews using complemented Baive Bayesian Classifier," Prithyi Information Solution Limited: India, International Journal of Intelligent Computing Research (IJICR), Volume 1, Issue 4, December 2010

[3] Liu B, "sentiment analysis and subjectivity, Hanbook of natural language processing," vol 2, pp 627-666, 2010

[4] S Padmaja1 and Prof. S Sameen Fatima, "Opinion Mining and Sentiment Analysis –An Assessment of Peoples' Belief: A Survey, International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1, February 2013

[5] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford. [6] Mitchell T 1997 Machine Learning. McGraw-Hill.

[6] Laporan Akuntabilitas Kinerja Tahun 2014 , Kementrian Pendidikan dan Kebudayaan, 2014

[7] Franky, Ondřej B,Veselovská K, "Resources for Indonesian Sentiment Analysis", The Prague Bulletin of Mathematical Linguistics, No 103, pages 21–41, 2015.

[8] Liu, B. Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.

[9] Steenkamp and Wedel, A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation. Journal of Marketing Research, pages 385-397, 1991.

[10] Hoda Waguih. A Data Mining Approach for the Detection of Denial of Service Attack. IAES International Journal of Artificial Intelligence, 2(2):99-106, 2013.