Framework of Sentiment annotation for document specification in Indonesian language Base on topic modeling and machine learning

Tata Sutabri
Faculty of Computer Science and Information Technology
Respati Indonesia University
Jakarta, Indonesia
tata.sutabri@gmail.com

Miftah Ardiansyah
Faculty of Computer Science and Information Technology
Gunadarma University
Jakarta, Indonesia
miftah.ardiansyah@gmail.com

Abstract-Reservation service users and or purchase online based on the marketplace often face difficulties in determining the object or service selected closest to the criteria of potential users. Aside from the rating or rating which features conventional, potential customers can make decisions with the customer review feature that has to wear or purchase items or services. The availability of these features provide a new task for prospective customers to get a thorough analysis, prospective customers are advised to read and analyze each comment related to the amount not less diverse language and style of Indonesian. The difficulty will be growing and time-consuming for prospective users when there are objects or services that are the same in different online services. This study proposes a framework to overcome the difficulties prospective customers. This framework implements a blend of approaches topic models, machine learning to perform sentiment analysis on services and purchase of objects or services based on online. The proposed framework has relevance or context of user reviews. Outcome future of this framework, including the form of the model ranking or rating based every existing review; due to the nature of the framework offered is specific to have a specific domain which minimizes missing review.

Keywords: Review online, framework, machine learning.

I. INTRODUCTION

Development of Information Technology, from time to time is very fast, so its role in human life can be felt in a variety of human activities, both individual, group, organization or company. Generally there are two types of textual information on the web that is fact and opinion. Facts are objective statements regarding activities, entities and events in human life, while opinions are subjective statements that reflect the sentiment or perception of people regarding activities, entities or events in human life.

When an organization or a company or a group or an individual, want to obtain public opinion on the image and services of a hotel, then they do not need to conduct a survey and focus group conventional expensive. Website as an online review sites, personal blogs, and social media provide sources of opinions in number for the needs of individuals and organizations. Through the website people can express anything, including his opinion of a thing without any compulsion. Use of the Internet is highly personal in every user, supported and sophistication of the technology allows the public gives a personal touch or "flavor" in real-time to almost all the online activities.

Portfolio services such as sites agoda.com, traveloka.com, tiket.com, booking.com, pegi-pegi.com and misteraladin.com allows users to submit reviews in the form of a short message or review, as customers or their customer. Thousands of customers are excited to use the existing facilities at the site services. They express their opinions, describe their experiences, and disseminate information about the quality of services provided, thus posting a review or a portfolio can potentially negatively or positively, on the service. One of the challenges that need to be solved is to identify a portfolio of some of the reviews or related to the same topic.

Challenges in data processing, which has a "sense" that led to the idea and implementation of research, one of which by analysis "taste", which is commonly called sentiment analysis. Not easy to measure the value of the sentiment of textual documents. Benefit analysis of sentiment in the business world, among others to monitor the service. The fast can be used as a tool to see the public response to the service. So it can be taken the next strategic steps. It encourages the need to do research sentiment analysis to review the document in Indonesian language. This research was conducted using the approach of models and machine learning topic devoted to the Indonesian-language text documents.

II. LITERATURE RIVIEW

Sentiment analysis or opinion mining is the process of understanding, extract and process the textual data automatically to get the sentiment of information contained in an opinion sentence. Sentiment analysis is done to see opinions or opinions tendency towards a problem or an object by a person, whether negative or positive tend to be opinionated. Sentiment Analysis is a branch of research in the domain Text Mining boom that began in the early 2002's. His research began to flourish since the paper of B. Pang and L. Lee out. In general, Sentiment analysis is divided into two major categories:

1) Coarse-grained sentiment analysis
2) fined-grained sentiment analysis

Coarse-grained sentiment analysis - the process of analysis at the level of documents. In a nutshell is the orientation classifying whole of document. This orientation there are three types: Positive, Neutral, Negative. However, there is also what makes the value of this orientation is continuous or discrete.

Fined-grained sentiment analysis - this second category is on the rise now. The point is that most researchers focus on this type. The object to be classified is not at the level of the document but a sentence in a document. Until now, most of the research in the field of sentiment analysis is intended only for English because tools or resources for the English very much at all. Some resources are often used for sentiment analysis is Word Net Sentiment and Word Net.

2.1 Topic Model.

Topic model is an algorithm that reveals hidden thematic structure in the document collection. Topics Model provides a simple way to analyze large volumes of text labeled. A "topic" consists of a group of words that often occur simultaneously. Using contextual clues, topic models can connect the word with the same meaning, and distinguish between the use of words with multiple meanings or meanings.

Modeling is a form of text mining topics, how to identify patterns in the corpus. You take your corpus and run it through a tool that groups of words in the corpus into 'topics. Miriam Posner has described modeling topics, as "method to find and track large groups of words within the structure of the text".

According to Brob, J., 2013, topic modeling approach is an effective tool to find the main theme in a set of customer reviews. It is also that the representation of the sentence document is more suitable in the context as well as setting the number of topics, between 50 and 70 is the most reasonable. In machine learning and natural language processing, the topic model is a statistical model to find the kind of abstract "topics" that occur in the document collection. Intuitively, given that the document is to have a particular topic, one would expect a certain word appears in a document more or less frequently. Output modeling this topic is not fully human readable. One way to understand this program is through visualization.

One way to think about how the process works is to imagine a topic modeling work through an article with a set of highlighter. When reading the article, using a different color for key words from the theme of the paper. When finished, the words are copied as a group of words. List of words is the topic, and each color represents a different topic.

To answer the problems encountered, necessary to formulate a strategy in the form of stages and steps in architecture. Topics approach used for the model can generate related topics in document or corpus. Topics to be the candidate best weights of the keywords of a theme or label. Annotated corpus compiled by implementing strategies generating a topic that is Latent Dirichlet Allocation (LDA).

2.2 Corpus.

Terminology dictionary and the corpus are sometimes interpreted or understood a similar though not the same. Often used interchangeably between the two terms. Both are in the form of a list of words. But the difference is, the dictionary lists words that more than corpus. Definition corpus according to John Sinclair is a collection or a collection of small parts or variations of a language, which is selected based on external criteria that represent a language or variety or variation of the language as a source of data for linguistic research.

The corpus contains a list of specific words in the subject area or a more specialized, such as business or economic corpus, the corpus financial, cultural corpus (Sundanese language corpus). The corpus can be understood as a source for the preparation of a dictionary, the dictionary has been compiled from previous corpus, as described in the historical development of the corpus. In some studies, the corpus is sometimes also called the corpora (plural of corpus). In addition corpus has an important role in a variety of research, especially related to language and linguistics, and various other researches.

Based on the general scope, the corpus is divided into two types:
1) Specific domain corpus, the corpus for the text to a special genre, for example, academic articles, business letters, and newspaper articles.
2) A common corpus that consists of text with a lot of different genres. Most studies conducted related to the corpus, using English as the language domain.

In this section, the authors register several corpus developed for English and non-English languages, including Indonesian language corpus of the target language in this writing.

2.3 Specific corpus.

The corpus according to the language used, divided into two categories, namely English and non-English. Some of the authors registered in available until today. For very much English corpus and related research has been done, as the initial research into the historical development of computational linguistics, corpus of the English language is dominant and widely developed.

As for the non-English corpus or minority language-related research is still growing, though not as advanced as the developments in the corpus of English. In the following explanation, we list some of the research corpus that is dominant in the development and use.

1). English Corpus

a) Princeton WordNet (PWN) or so-called WordNet, which is claimed as the corpus or lexical database of English-speaking world built by the University of Princeton (http://wordnet.princeton.edu/wordnet/). The surplus is accessible for free by anyone by providing software that is ready to be downloaded and installed by the user, or can use it via online (http://wordnetweb.princeton.edu/perl/webwn). But unfortunately, it does not support WordNet Indonesian.

b) WordBanks Online or Online Wordbanks Collins (http://www.collins.co.uk/page/Wordbanks + Online), which is based on English corpus comprising 550 million words are constructed from a variety of sources, both written and oral, and contains 8 varieties of English. The majority of the source taken in the period 2001-2005. As of this writing quoted, unfortunately the development process corpora non-English language is under development. To gain access to these corpus users must subscribe for a nominal amount.

c) ICE or the International Corpus of English (http://ice-corpora.net/ice/), a collection of corpora representing varieties of English from around the world. More than 20 countries that use English as a first language or English as the official language to the two countries. Indonesia was not included in the list of countries involved in this corpus, the closest neighboring countries involved are Singapore, Malaysia, the Philippines and Australia (http://en.wikipedia.org/wiki/International_Corpus_of_English).

d) British National Corpus (BNC), Developed by the University of Lancaster and Oxford University Computing services with a collection of 100 million words of English (British) well-written text or speech. BNC was built as the representative of the standard language in the general sense, is not limited to regional variations on the type of text or narrow (ex. scientific texts, writing papers, etc.).

2) Non-English Corpus

Non-English corpus is a corpus used other English as the target language. The development of the corpus of this category increased from time to time in line with their respective needs in each country, including: The corpus of the Spanish language; Arabic language corpus; Norwegian language corpus; Finnish language corpus; Chinese corpus; Online corpus in Filipino as it was built as part of an online repository Palito. The corpus is inspired and refers to the International Corpus of English (ICE); Thailand language corpus, ex. LOTUS (Large vocabulary Thai Continuous Speech Corpus); corpus Japanese, Japanese WordNet corpus gloss (JSEMCOR corpus) the adjustment corpus of English and Japanese.

Indonesian corpus luteum While the majority of this category using the corpus of English such as WordNet, ICE and others as essentially corpus which is then translated into the target language of each country.

2.4 Semantic Annotation

Annotations or labeling is to add linguistic information into a corpus. For example, the type of annotation is usually done by adding a tag or label to indicate that a class of words from a text. Annotations are commonly referred to part-of-speech tagging (POS tagging), which is useful, for example, to distinguish between words that have the same pronunciation or spelling or pronunciation but different meanings. For example, the noun (n) "could be" significant toxic substances or something bad and a verb (v) "could" be meaningful or capable; or the noun "city" means a berth with the noun "city" means ditches or trenches and the noun "city" as a person who controls the action either hidden or not.

2.5 Machine Learning.

1) Supervised Methods.
In order to train the classification for sentiment in text recognition, classification supervised by learning techniques, such as Support Vector Machines, Multinomial Naive Bayes and Maximum Entropy. Supervised approaches require the use of labeled training corpus to study the function of specific classification. The method often produces the highest accuracy, is using Support Vector Machine classifier.

a) Support Vector Machines (SVM)
Support Vector Machines, operates by building hyperplane with maximum Euclidean distance at closest training examples. It can be seen as the distance between hyperplane separator and two hyperplanes parallel on each side, representing the limit of examples of the class in feature space. It is assumed that the best generalization of this classifier is obtained when the maximum distance. If the data is not separated, hyperplane will be selected who shared data with minimal errors as possible.

b) Naive Bayes Multinomial (NBM)
A Naïve Bayes, using Bayes' rule (which states how to update or revise believe in the light of new evidence) as the primary equation, with the naive assumption conditional: each individual feature is assumed to be an indication of the class, not tied to one another. A multinomial classification builds the model by fitting the distribution of the number of occurrences of each feature for all the documents.

c) Maximum Entropy (MAXENT)
The approach tries to preserve as much uncertainty as possible. Some models are calculated, in which each feature according to the constraints on the model. Models with maximum entropy models that satisfy all constraints have to be classified. This way there is no assumption is made that is not justified by empirical evidence that there is.

2) Unsupervised and Weakly-supervised Methods.
The above techniques all require labeled corpus to learn classifiers. It is not always available, and it takes time to label a corpus of significant size. Method unsupervised can label a corpus, which is then used for unsupervised learning (especially semantic orientation help for this). Turney technique using AltaVista can be seen as a form of weakly supervised learning, where one set of terms seed extended to the collection of words.

There are two more methods to determine the sentiment of single words based on weak-supervised method. Hatzivassiloglou and McKeown presented a method to determine the sentiment of adjectives by classifying documents into equal parts-oriented, and manually labeled positive or negative cluster. A system that uses the term semantic clustering to determine the orientation of said opinion in combination with other words in the sentence. The idea behind this approach stems from the fact that the orientation word can change with respect to the features related word or phrase (ex. hot word in the pair: the hot water has a positive sentiment, but in a couple of hot space have a negative sentiment).

III. STATE OF THE ART

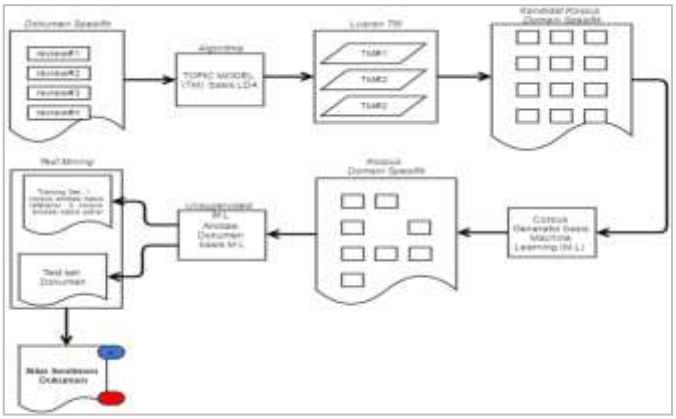
Related research and support in the development of the framework for specific document annotations sentiment Indonesian language based topic models and machine learning, are:

No	Type Publication	Author & YEAR	Topic Research	Novelty
1	Journal	Jo, Y., dan Oh, Alice, 2011.	Aspect and Sentiment Unification Model for Online Review Analysis. Dataset: electronic and restaurant	The combination of S-LDA method and Asum.
2.	Presiding	Boiy, E.; Hens, P.; Deschacht, K.; Moens, M.F., 2007	Automatic Sentiment Analysis in On-line Text. Dataset: review film	Test techniques sentiment analysis. Tools SVM and Naive Bayes Weka (software) Max Entropi Open NLP (software).
3	Dissertation	Brob, J., 2013	Aspect Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques	Distant supervision techniques to reduce human supervision in the process of annotation (automatic machine learning). Accuracy 91% correct in labeling the viewing process corpus.
4	Journal	Rui Xie, Chunping Li, and Qiang Ding, Li Li, 2014	Integrating Topic, Sentiment and Syntax for Modeling Online Review	Combines Post (Part of Speech) in the model. Tag Sentiment Aspect Models (TSA).
5	Journal	Ivan Titov, Ryan McDonald, 2008	Modeling Online Reviews with Multi-grain Topic Models	Unsupervised Topic Model Multi-grain LDA.
6	Journal ACM	Changlin Ma, Meng Wang, dan Xuewen Chen, 2015	Topic and Sentiment Unification Maximum Entropy Model for Online Review Analysis. Objek: restaurant	Sentiment classification with Maximun Entropy. Topic Sentiment Unification methodology with Maximum Entropy-LDA (TSU MAXENT-LDA). Better than the model (62.265): Asum and Sentiment-LDA.
7	Journal	Raja Mohana S.P, Umamaheswari K, dan Karthiga R, 2015	Sentiment Classification based on Latent Dirichlet Allocation	SVM for classification sentiment / opinion CHI value calculation
8	Journal IEEE	Bin Lu, Myle Ott, Claire Cardie, dan Benjamin K. Tsou, 2011	Multi-aspect Sentiment Analysis with Topic Models	Multi-aspect topic model *(weakly supervised TM).* The rating sentiments provide a document
9	Journal ACM	Neil O'Hare , Michael Davy, A.Bermingham, Paul F., Páraic S, Cathal G, Alan F. Smeaton, 2009	Topic Dependent Sentiment Analysis of Financial Blogs. Data: blog financial	Supervised (machine) learning. Classification sentiment with Multinomial Naive Bayes classifier

10	Presiding	M. Yusuf Nur, Diaz D. Santika, 2011	Analyst sentiment on Indonesian Language Approach Document Support Vector Machine	Classification of documents by language use SVM approach 73.07% accuracy rate
11	Dissertasion	Miftah Ardiansyah 2015	Annotated corpus-based topic models for client analysis system supporting consular dialogue	Topic models can predict the topic of a document. Building a prototype system supporting analysis of CCD-based topic models
12	Presiding	Ghulam Asrofi Buntoro, Teguh Bharata Adji, Adhistya Erna P. 2014	Twitter Sentiment Analysis with Combination Lexicon Based and Double Propagation	Generate 7 parameters sentiment analysis, with accuracy of 23.43%.
13	Journal	Raisa Sianipar, Erwin Budi Setiawan 2015	Strength detection sentiment on Indonesian language Tweet text Using Sentistrength	Build a system that adapts SentiStrength classifier with accuracy grades of 57.33%

IV. METHODOLOGY

Below is the methodology used in this study is:



The purpose of this study was to classify documents based on the language to the positive or negative sentiment. The methodology starts from the process of data collection and extraction, selection and generate word corpus, as well as process and document annotation clustering. Corpus of data used in this study was obtained from the document review about the image and services of a hotel, on the site agoda.com, traveloka.com, tiket.com, booking.com, pegi-pegi.com and misteraladin.com

3.1 Data collection and extraction.

Data used in the form of online reviews, which is obtained from the site agoda.com, traveloka.com, tiket.com, booking.com, pegi-pegi.com and misteraladin.com, by adding detection process language to get the data in Indonesian language. Topics are limited regarding the image and services of a hotel review. Document review that has been obtained from multiple sites, then extracted using the topic model, to produce multiple outputs simultaneously be candidates for a specific domain corpus, as the first training set.

3.2 Selection and generate word corpus.

After candidate corpus or training set is first formed, it is necessary to make the selection of words to be selected or automatically selected to generate specific domain corpus using machine learning, a training set of the second. Steps being taken are as follows:

1) Cleansing, namely the process of cleaning the document of words that are not required to reduce noise. The word omitted is the HTML code, keyword, emotion icons, hash tag (#), username (@ username), url (http://situs.com), and email (nama@situs.com).
2) Case folding, namely uniformity of shape as well as the elimination of the numbers and punctuation. In this case the used only latin letters between a through z.
3) Parsing, which is the process of breaking into a word document.

3.3 Document annotation and clustering

The next stage, after the candidate corpus or training of the second set is formed, then the next step, perform annotation of documents using machine learning next, to produce training third set form of corpus annotation-based reference (dictionary), and corpus annotation-based experts, and test sets of documents.

After the third training set is formed, the next process is the process of annotation using machine learning by using one of the methods of clustering to generate positive and negative sentiment perspective of the documents that have been extracted. Steps being taken to the clustering process are as follows:

1) Part of Speech (POS) tagger,
namely the process of giving a class on the word. Class selected word is an adjective, adverb, a noun and a verb that the four types of words above are the kind of words that most contain sentiment. Determination of the class of words by big Indonesian dictionary.

2) *Stemming*, aims to reduce variations of words
that have the same basic words. As in the POS process Tagger, stemming process is done with the help of big Indonesian dictionary.

V. CONCLUSION

From the results of this study concluded that the framework for a specific document annotations sentiment Indonesian language is a new model that has a level of accuracy that is better than the previous study. This research is expected to contribute as follows:

1) Help give sentiment or "flavor" of a collection of textual documents
2) Assist in injecting a semantic intent terminology in sentiment analysis
3) Being specific corpus for related research, particularly in any comments or textual review,
4) Providing added value of each activity ranking (rating) a review.

REFERENCES.

[1] Boiy, E.; Hens, P., Deschacht, K., Moens, M.F., Automatic Sentiment Analysis in On-line Text. Proceeding, 2007.

[2] Brob, J., Aspect Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques, Dissertation, 2013.

[3] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou., Multi-aspect Sentiment Analysis with Topic Models,. Journal IEEE, 2011.

[4] Card S.K., Mackinlay J.D. Shneiderman B. (eds), 2009, Reading in Information Visualization, Using Vision to Think, San Francisco: Morgan Kaufmann

[5] Changlin Ma, Meng Wang, and Xuewen Chen., Topic and Sentiment Unification Maximum Entropy Model for Online Review Analysis., Journal ACM, 2015.

[6] Dumbill, E. 2012. *Big Data Now Current Perspective.* O'Reilly Media.

[7] Eaton, C., Dirk, D., Tom, D., George, L., & Paul, Z. (n.d.). 2012. *Understanding Big Data.* Mc Graw Hill.

[8] Ivan Titov, Ryan McDonald., Modeling Online Reviews with Multi-grain Topic Models, 2008.

[9] Jo, Y., Oh, Alice, Aspect and Sentiment Unification Model for Online Review Analysis. Journal ACM, 2011.

[10] Miftah Ardiansyah, Annotated corpus-based topic models for client analysis system supporting consular dialogue. Dissertation, 2015.

[11] Neil O'Hare , Michael Davy, A.Bermingham, Paul F., Páraic Sheridan, Cathal Gurrin, Alan F. Smeaton., Topic Dependent Sentiment Analysis of Financial Blogs. Journal ACM, 2009.

[12] Rui Xie, Chunping Li, dan Qiang Ding, Li Li., Integrating Topic, Sentiment and Syntax for Modeling Online Review. Journal, 2014.

[13] Raja Mohana S.P, Umamaheswari K, and Karthiga R., Sentiment Classification based on Latent Dirichlet Allocation., Journal, 2015.

[14] Sianipar Raisa and Budi Erwin Setiawan., Strength detection sentiment on Indonesian language Tweet text Using Sentistrength. Journal, 2015.

[15] V. Jijkoun, M. de Rijke, and W. Weerkamp. Generating focused topic-speci_c sentiment lexicons. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 585{594, Uppsala, Sweden, July 2010. Association for Computational Linguistics.