A Hybrid Model for Social Media Sentiment Analysis for Indonesian Text

Syopiansyah Jaya Putra[1], Ismail Khalil[2], Muhamad Nur Gunawan[3], Riva'I Amin[4], Tata Sutabri [5]
[1,3,4]Faculty of Science and Technology, Syarif Hidayatullah State Islamic University Jakarta
[2]Institute of Telecooperation, Johannes Kepler University Linz, Austria
[5]Faculty of Information Technology, Respati University of Indonesia
syopian@uinjkt.ac.id[1], ismail.khalil@jku.at[2], nur.gunawan@uinjkt.ac.id[3],
rivai.amin@mhs.uinjkt.ac.id[4] , tata.sutrabri@gmail.com[5]

ABSTRACT

Sentiment analysis for Indonesian social media text is very important because the text content in social media is very diverse and requires an accurate method that can produce an analysis describing the state of the actual data. The main problem in sentiment analysis for Indonesian text on social media is unstructured text data and the use of non-standard languages such that sentiment analysis often produces errors. This paper focuses on sentiment analysis using a hybrid model that combines lexicon based and maximum entropy methods to classify the sentiments of Indonesian public opinion on government. The method consists of extracting datasets, preprocessing, lexicon-based classification, machine learning training, machine learning classification, and result interpretation. The results of the study produce 91 classifications of neutral sentiment, 51 document negative sentiments, 39 document positive sentiments and 152 document of mix sentiments. Based on the evaluation results, the hybrid sentiment model for Indonesian Language sentiment analysis on social media produced a pretty good accuracy score of 84.31% compared to previous studies. The implication of this study is to produce a sentiment analysis system with a hybrid method for Indonesian text on social media.

CCS Concepts

- **Information systems → Information retrieval → Retrieval tasks and goals → Sentiment analysis**

Keywords

Sentiment analysis, hybrid based, lexicon based, maximum entropy, sentiment classification.

1. INTRODUCTION

Sentiment Analysis of social media is essential for analyzing the opinions, sentiments, evaluations, judgments, attitudes and emotions of people towards certain entities, such as products, services, organizations, individuals, issues, events, topics and related attributes [1].

Research in the field of sentiment analysis began in 2002. Turney [2] conducted a research with the sentiment theme of a consumer review analysis of a product. The method in [2] is Pointwise Mutual Information (PMI) for the estimation of Semantic Orientation (Semantic Orientation) a phrase, with motor vehicle review data accuracy reaching 84% and film review data by 66%.

Pang et.al. [3] classify film reviews at document level that have positive or negative opinions using supervised learning techniques. Collection of movie reviews that have been determined positive or negative are used as training data for some machine learning algorithms. Accuracy obtained in the study [3] ranged from 72% to 83%.

Based on the design of Medhat et al [4], the method in sentiment analysis is divided into two approaches, the first is machine learning and the other is Lexicon-based [4]. Machine learning is divided into two main categories: supervised and unsupervised. The success of both is based on the selection and extraction of suitable features, for example: (1) terms (words or n-grams) and their frequency; (2) part-of-speech information; (3) the word negation in each sentence; (4) syntax dependence (tree parsing). While lexicon-based or vocabulary-based approaches rely on sentiment vocabulary, which is a collection or collection of words, phrases and sentiment idioms that are known.

Supervised learning provides more accurate results but making training data to build sentiment models requires large amounts of data [5]. In previous studies, training data was often made manually or using data that had been defined in advance. However, Putranti and Winarko [6] conducted a study based on lexicon which produced automatic training data of 44,006 documents and 18,069 words. This study has an accuracy value of 86.81%.

Hybrid-base is a combination of Lexicon-based and machine learning. The Lexicon-based classifier has the duty to generate data that has been labeled, the data is used as training data to produce a sentiment model using a machine learning classifier. Furthermore, the Lexicon-trained learning-based hybrid classifier is run to test test data [7].

Maximum entropy is a framework for integrating information from a variety of different sources for classification purposes. Data for classification is made of a number of features that are slightly complex and allow researchers to use existing knowledge to determine the type of information that is important for classification [8].

iiWAS '18, November 19–21, 2018, Yogyakarta, Indonesia

ACM ISBN 978-1-4503-6479-9/18/11…$15.00
https://doi.org/10.1145/3282373.3282850

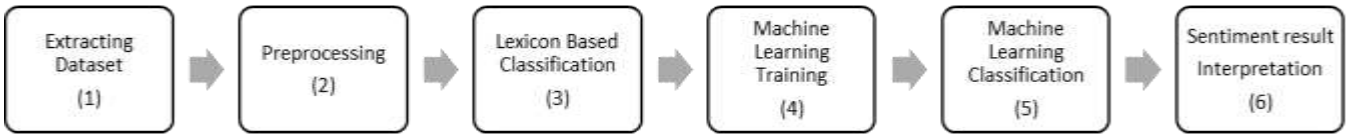

Figure 1. A Hybrid model for Sentiment analysis

In a government perspective, designing and implementing policies at any level in government is a complex process. One of the difficulties is finding and summarizing public sentiments and opinions. Communities are not too active in participating in e-Government portals and policy specialists lack equipment to take into account people's views on policy issues expressed in real time through social networks. The main challenge of this condition is to assess the characteristics of users who discuss government issues on social media and assess whether the user truly represents public opinion [9].

This paper aims to analyze text sentiment on Indonesian social media by combining lexicon based and maximum entropy methods to produce a sentiment analysis system that can be used to classify sentiments from public opinion on the government. This study uses a dataset of 6000 texts from social media. Methods are extracts of datasets, preprocessing, lexicon-based classifications, training datasets, and machine learning classification. This study produces 333 document sentiment classification which consisting of neutral sentiment 91 documents, negative sentiment 51 documents, positive sentiment of 39 documents, and mixed sentiment of 152 mixtures.

Based on the evaluation results, a hybrid sentiment model for the analysis of Indonesian sentiment on social media resulting in an accuracy value of 84.31%, this result is quite good compared to similar studies on Indonesian language datasets.

In the second section describes the methodology, section three discusses the results of the experiment, and the final part is about the discussion on the application of the hybrid method.

2. RESEARCH METHOD

Sentiment analysis methodology consists of preprocessing stage, lexicon-based classification, machine learning and machine learning classification training such as in Figure 1.

Dataset retrieved from data extraction to use Application Programming Interface (API) comprising 6951 Twitter tweets. The Dataset is obtained by using the Twitter API and tool named twurl to pull (crawling) tweet data that relate to the keyword "Government", the result of the crawling in the form of a format JavaScript Object Notation (JSON). The data is then filtered only allows the general public to be used, so the official account (verified) was not taken in the include. Then data included must be an original tweet so that the data in the form of a retweet should be discarded. And the last data is extracted just taken field will be needed, such as the ' id ', ' text ', ' user_id ', ' user_name ', ' created_at ', ' source '.

2.1 Preprocessing

This step consists of cleansing, tokenisasi, case-folding, stop word removal and stemming. Stages of cleansing, namely cleansing Word – the word that is not necessary as the hashtag, URL, mention, and duplication of data. Later phases of the tokenisasi [10] i.e. separating words based on whitespace, then case-folding with changing text from capital becomes a non-capital or small-case, next the stop word removal by removing tokens that are not required in accordance with the stopword list, and last words or vocabulary is converted back into basic shapes by removing the prefix.

2.2 Lexicon Based Classification

This stage consists of labeling the dataset to the specified preprocessing results where words containing positive and negative sentiment lexicon, lexicons also negation using dataset sentiment lexicon derived from previous research [11, 12]. Then proceed with the classification based on the lexicon with the following conditions:

a. If it contains words with positive sentiment is more positive than negative so categorized with the label "10";

b. If it contains words with negative sentiment is more positive than negative with categorized then label "01";

c. If the amount of positive and negative sentiment then categorized the mixture with the label "11";

d. and if not have both categorized neutral with the label "00".

After it's done expert validation to verify whether each tweet's been labeled correctly, if the sentiment is still less precise then the lexicon sentiment need to be adjusted and then repeated again this stage.

2.3 Machine Learning Training

On the process of machine learning training using OpenNLP Document Categorizer Sentiment to train the Model. This process covers three stages: application to load the file Training Data; then call DoccatTrainer commands, this is the Maximum Entropy framework used, this process will result in a binary format with the model file.

2.4 Machine Learning Classification

This stage is the core process of the analysis of the test data, where sentiment already separated will be classified using the resulting sentiment model at this stage of training. This stage uses a web-based application, with the Apache Solr as the database server to be injected from the OpenNLP library.

This stage using OpenNLP Tools by means of (1) the resulting sentiment model Files at this stage of Training are loaded on DocumentCategorizerME. (2) any resulting sentence at pre-launch stage processing used to categorize DocumentCategorizerME. determine the sentiment of what is contained, and (3) Called DocumentCategorizerME function getBestCategory to determine the category. the sentiment of most approach, (4) by the Apache Solr, data analysis results sentiment will be stored along with the attributes id, created_at full_text, and sentiment.

2.5 Evaluation

The evaluation is carried out by comparing the file data with sentiment training models that have been created, using OpenNLP tools DoccatEvaluator. The results obtained are in the form of

score accuracy shows the accuracy of the model of sentiment. As for the calculation for the value of accuracy can be done in the following way [13-18]:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{sum\ of\ all\ predictions} \quad (1)$$

The number of correct predictions was the result of a precise analysis of the sentiment fits the content of the training data, while the number of all the predictions were the result of sentiment analysis.

3. RESULT AND DISCUSSION

The result of the method of hybrid that is implemented in the analysis of social media sentiment on the Indonesian Language with the keyword "Government" is the classification of text documents of twitter.

3.1 Preprocessing

Results of crawling twitter data with the keyword "Government" is the JSON format with the file, the result of the crawling managed to get the tweet data as much as 7658.

Based on the results, the data preprocessing tweet was successfully reduced to 6951 and has also been cleared of hashtag, mention, and URL. As well as have also been converted into lower case, omitted stopword and converted into Word.

3.2 Lexicon Based Classification

In this process, there are three processes that are performed, i.e. Lexicon labeling, lexicon based classifier, and expert validation.

a. Lexicon Labeling

Found with a total of 2758 negation words, words with positive sentiment as much as 21070 and words with negative sentiment as much as 21424. The following is an example of the results of tagging the word with the negative and positive sentiment, and negation.

Results from the process of labeling the lexicon sentiment and the negation. It can be seen for instance in the tweet :"Your campaign is out of date the boss, not creative at all, just brought public opinion in order to hate with the Government now ... shame on you ..", the word "hate" is labeled negative sentiment that it became "hate"|negative also with the word" good "are labeled with the positive sentiment that it becomes" good "|positive. There is also a tweet that does not experience the labeling at all because of undetected said sentiment or negation.

b. Lexicon Based Classifier

From the tagging can produce a corpus with the number of Tweets containing positive sentiment as much as negative sentiment as much as 2464, 1821, the mixed sentiment as much as 806 and neutral sentiment as much as 1860.

On the tweet starts with a binary code "00" which means that the tweet is neutral, then classified on tweet code beginning with "10" meaning the positive sentiment classified tweet, tweet later begins with code "01" which means that the tweet is classified as a negative, and the last the tweet is started with code 11 which means the tweet classified sentiments mixed.

c. Expert Validation

With expert validation, there are changes in the dataset is lexicon, that is the negative of the lexicon into a number of 2309 lexicon become positive, a number of 1159 lexicon and lexicon negation into a number of 15 lexicon.

For example on the tweet is shown that neutral sentiment, even though there is a word "on", "dumb" and "emotion" that is negative, so the lexicon should tweet is the negative sentiment. The negative words should be included in the database of the lexicon and then recount done.

3.3 Machine Learning Training

Based on the training process using OpenNLP DoccatTrainer, there are countless 6951 events on the training data, then disposed of some of the events so that it becomes the event 6857 after it is compressed and merged so that it becomes a number 6679 event. Then on the stage of indexing data training, produce outcomes a number 4 and number of predicate 2264. After that held as many as 100 models sentiment training time looping. Model sentiment then saved on file "id-sentiment. bin".

3.4 Machine Learning Classification

Based on the classification process with machine learning using a model previously generated sentiments, then those documents have been identified as many as 333 sentiment classification of documents with details of neutral sentiment as much as 91 document, document 51 negative sentiment, sentiment was positive, and the documents 39 mixed sentiment as much as 152 documents.

3.5 Evaluation

Based on the evaluation of the process, from the training data, 12481 with average calculation as much as 59433.3 documents per second, 0.21 seconds counting long, and long process for 0384 seconds, the value of accuracy obtained is 0.8431 or 84.31 %.

Based on the research results, approaches and lexicon-based machine learning algorithm with maximum entropy by using the dataset Indonesia-speaking opinion especially regarding Governments can generate model sentiment value accuracy about 84.31%. The results obtained by using the training data as much as 6951 tweets.

3.6 Discussion

This study is already producing a neutral sentiment as much as 91 documents, sentiment negative 51 documents, document 39, positive sentiment, and mixed sentiment as much as 152 documents with 84.31% accuracy rating. When compared to the previous research [5] which also uses a hybrid approach, the value of the resulting accuracy on this research is still under research and Putranti Winarko [5]. Based on the results of such research, SVM yields a value of highest accuracy with 86.81%, with Maximum Entropy assisted as POS Tagger.

The possibility of such differences are caused by the difference in the training data used IE as much as 44,006 documents. Because according to the theory, in supervised learning, the more training data will be the better the results. However, a number of time-consuming processes as much as 1688 seconds, a difference much with this research which takes 2.7 seconds to train the model.

According to research conducted by Bhatt et al. [19] sentiment analysis, yielding the value of better accuracy if combining lexicon based and machine learning. Combines lexicon based and machine learning is not the first time this is done, the use of the lexicon-based and machine learning simultaneously applied on

such research that is also using maximum entropy value accuracy can reach 80%. It is also shown that the more data that is used in the training data will also be getting great results.

When compared with previous studies which also using maximum entropy, for example, research by Lunando and Purwarianti [15] produces 78.4% accuracy for classification and labeling directly 980 training data manually. Later on research results Manurung [20], the best accuracy value for the translation of Indonesia was 80.46% using LastInclude Language Data and Maximum Entropy. Then different results are also given by Naradhipa research and Purwarianti [21], using the dictionary for the extraction of features and convert the text into a formal form of preprocessing stage is applied to the data, has been training 150 produce a value of 80% accuracy if using maximum entropy. Whereas, in this study, the value of the fit reaches 84.31% of the authors can conclude it can prove the theory advanced by Patel et al.

However, different results shown if using different languages, for example for classification of sentiment against the United Kingdom language film reviews done by Mehra et al. [22], can generate value accuracy reaches 85% with the unsupervised use of maximum entropy. Then on the classification of Chinese-speaking products reviews conducted by Lee and Renganathan [23] yields a value of 87% accuracy. As for the classification of the Spain-language tweet using the data produce a value of 78% accuracy. If viewed from the value of the resulting accuracy, the use of maximum entropy language against the possibility of Indonesia can be increased again.

Research conducted by the author using the combined approach between lexicon based and supervised maximum entropy. Based on author searches, there is still no direct mention research using a combination of both methods on the analyses of Indonesia speaking sentiment, most application approaches was applied in the research of Putranti and Winarko [6].

With his model system analysis of the sentiment and sentiment to public opinion, in general benefits that can be drawn is that we or a user can classify sentiment that originally could not be identified directly and automatically on the Tweet or text, can be directly identified directly, automatically and quickly.

In addition with the formation model of sentiment to public opinion against the Government, the implementation of e-government can be supported, as Liu et al. [9] mentioned above that find and sum up the sentiment and public opinion is one of the key difficulties. With the existence of a system of analysis of sentiment, society as involuntarily actively participated in the implementation of e-government, and the Government can take into account the views of the community against the policy issues that are expressed through social networking.

Then successfully applied to the analysis of sentiment by using the lexicon based and supervised machine learning at the same time, this research can be used as a reference for other researchers are also interested in conducting research on sentiment analysis. Remember this is still rarely used methods and a lot of things that can be explored.

Other benefits obtained from the results of this research is to provide an alternative method in shaping the training data used to train the model. Because one of the difficulties in applying machine learning by supervised learning approach is that we must make the data work with large numbers [5], and some still do that by labeling one by one on the training data. In this study, a trainer can form data automatically, and a little effort to completion.

4. CONCLUSION

This research focuses on the creation of a model for sentiment analysis using a combination of a lexicon based and supervised maximum entropy, with stages of the training, classification, and evaluation, for classifying sentiments to be positive, negative, neutral and mix in public opinion about the Government.

This study produces a neutral sentiment as much as 91 classification documents, sentiment negative 51 documents, document 39, positive sentiment, and mixed sentiment as much as 152 documents. Based on the evaluation results, the model for the analysis of hybrid sentiment Indonesian Language in social media generates a pretty good accuracy value i.e. of 84.31% compared to previous studies.

This research resulted in the model of sentiment that can be used for classifying public opinion against the Government became sentiment positive, negative, neutral and blends. The system also produces a sentiment analysis can index and sentiment analysis results searching data analysis results tweet sentiment. Based on the results of the evaluation, models of accuracy providing value sentiment 84.31%.

The implication of this study is to help the user to establish the training data automatically, classifying sentiment in hybrid, and taking into account the views of the community against the policy issues that are expressed through social networking.

5. REFERENCES

[1] B. Liu, "Sentiment Analysis and Opinion Mining," p. 168, Apr. 2012.services," Information Sciences, vol. 311, pp. 18–38, Aug. 2015.

[2] P.D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics. 2002, July

[3] B. Pang, L. Lee, & S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics. 2002, July.

[4] W. Medhat, A. Hassan, & H. Korashy,Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113. 2014.

[5] S. B.,Kotsiantis, I. Zaharakis, & P. Pintelas. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24. 2007.

[6] N. D. Putranti and E. Winarko, "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine," p. 10, Jan. 2014.

[7] F. Sommar and M. Wielondek, Combining Lexicon- and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification. 2015.

[8] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," Journal of King Saud University - Engineering Sciences, Apr. 2016.

[9] M. S. Neethu, & R. Rajasree. Sentiment analysis in twitter using machine learning techniques. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-5). IEEE. 2013, July.

[10] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the Web," 2005, p. 342.

[11] M. Sokolova and G. Lapalme, (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

[12] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, & M. Stede. Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307. 2011.

[13] R. Feldman. Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89. 2013.

[14] D. Patel, S. Saxena, T. Verma, and P. G. Student, "Sentiment Analysis using Maximum Entropy Algorithm in Big Data," vol. 5, no. 5, p. 7, 2007.

[15] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," 2013, pp. 195–198.

[16] B. Pang, & L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics. (2004, July)

[17] Putra, S. J., Mantoro, T., & Gunawan, M. N. (2017, November). Text mining for Indonesian translation of the Quran: A systematic review. In Computing, Engineering, and Design (ICCED), 2017 International Conference on (pp. 1-5). IEEE.

[18] T. Sutabri, S.J. Putra, M.R. Effendi, M.N. Gunawan, D. Napitupulu. Sentiment Analysis for Popular e-travelinh Sites in Indonesia using Naïve Bayes. The 6th International Conference on Cyber and IT Service Management (CITSM 2018). 2018.

[19] A. Bhatt, A., A. Patel, H. Chheda, & K. Gawande. Amazon Review Classification and Sentiment Analysis. International Journal of Computer Science and Information Technologies, 6(6), 5107-5110. 2015.

[20] R. Manurung, "Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews," ResearchGate, 2008.

[21] A. R. Naradhipa and A. Purwarianti, "Sentiment Classification for Indonesian Message in Social Media," p. 4, 2011.

[22] N. Mehra, S. Khandelwal, and P. Patel, "Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews," p. 7, Jan. 2002.

[23] H. Y. Lee and H. Renganathan, "Chinese Sentiment Analysis Using Maximum Entropy," p. 5, 2011.