Available Online : https://proceeding.researchsynergypress.com/index.php/cset/index
RSF Conference Series: Engineering and Technology
ISSN 2809-6843 (Online) | 2809-6878 (Print)
Volume 1 Number 1 (2021): 364-367

Lexicon Based Sentiment Analysis in Indonesia Languages: A Systematic Literature Review

Yuli Fauziah [1], Bambang Yuwono [2], Agus Sasmito Aribowo [3]

[1] Lecturer from Information System Department, Universitas Pembangunan Nasional "Veteran" Yogyakarta;
[2] Lecturer from Informatics Department, Universitas Pembangunan Nasional "Veteran" Yogyakarta;
[3] Lecturer from Informatics Department, Universitas Pembangunan Nasional "Veteran" Yogyakarta. PhD Student at FTMK UTeM Malaysia

Abstract

This systematic literature review aims to determine the trend of lexicon based sentiment analysis research in Indonesian Language in the last two years. The focus of the study is on the understanding of preprocessing used in lexicon-based sentiment analysis studies in the last two years, the lexicon used in these studies, and classification accuracy. The main question in this SLR : what techniques of lexicon based sentiment analysis will provide the highest accuracy. The most widely used preprocessing methods in previous research are tokenization, case conversion, stemming, remove punctuation, remove stop word, remove or replace emoji and emoticons, and normalization or slangword conversion. The sentiment labeling process in previous studies calculated based on the comparison of the number of negative sentiment keywords with positive sentiment keywords in one sentence. The maximum accuracy from previous study is 90%. The most widely used lexicon is NRC and Inset which is a lexicon dictionary in Indonesian. Knowledge of this can be used to propose a better model for lexicon based sentiment analysis in Indonesian Languages.

Keywords: Lexicon, Sentiment Analysis, Indonesian Language, accuracy, preprocessing

INTRODUCTION

Sentiment analysis is still an interesting topic for research in the field of text mining. Sentiment analysis provides various implementations, especially it is also used to determine user ratings of certain products, political figures, services and brands. The development of social media has caused sentiment analysis to gain a place, because social media has become a tool for mass communication and opinion. There are billions of social media users, they can give their opinions and opinions freely, not limited by place and time. To interpret these opinions into positive and negative opinions, of course, tools and methods are needed. A good sentiment analysis method is a method that will guarantee the accuracy of sentiment according to actual conditions.

There are two approach paradigms to get sentiment from opinion, namely sentiment analysis based on lexicon and based on machine learning. Lexicon consists of two types, namely dictionary-based (Sasmito Aribowo, 2018) or corpus-based approaches. This means that the sentiment in a sentence is calculated based on certain computations based on a language dictionary that already contains knowledge. This method is growing rapidly because the use of dictionaries, especially digital Indonesian dictionaries, is also developing. There are several Indonesian dictionaries available, such as the NRC Lexicon which also contains keywords in Indonesian, and the Inset (Indonesian Sentiment Lexicon).

This systematic literature review aims to determine the trend of lexicon-based sentiment analysis research in Indonesian in the last two years. Why Indonesian? Indonesian has characteristics that do not exist in other types of language. These characteristics include word structure, grammar, and semantics. Another fact is that opinions in Indonesian contain a lot of stop words and slang words (Khomsah & Aribowo, 2020). Therefore, an in-depth study is needed to find the latest methods or the best models for sentiment analysis in lexicon-based Indonesian. The focus of the study is on the understanding of preprocessing used in lexicon-based sentiment analysis studies in the last two years, the lexicon used in these studies, and classification accuracy.

Corresponding author:
yuli.fauziah@upnyk.ac.id; bambangy@gmail.com; sasmito.skom@upnyk.ac.id
DOI: 10.31098/cset.v1i1.397

After knowing the trend of lexicon-based sentiment analysis research, it will be concluded that the preprocessing model is the most widely used, the best lexicon, and the highest accuracy. Knowledge of this can be used to propose a better model for sentiment analysis in Indonesian using the lexicon.

RESEARCH METHOD

a. SLR Question

The main question in this SLR: what techniques of lexicon-based sentiment analysis will provide the highest accuracy. To specify the main question, a list of short questions is defined in Table 1.

Table 1: SLR Question

No	Question	Motivation
Q1	What are the pre-processing techniques often used?	To know how significant the pre-processing stage to the accuracy of sentiment and emotion analysis.
Q2	What are the method used in labelling the sentiment ?	To categorize types of method used in sentiment emotion analysis.
Q3	Which lexicon are best for classifying twitter data?	To find out what lexicon give high accuracy
Q4	What is percentage of accuracy of sentiment emotion analysis?	To find out the classification method give highest accuracy.

b. Literature Search Strategy

We analysed selected articles that have been reviewed in peer-reviewed and published in many journals and proceddings. The search query used is:

“sentiment” AND “analysis” AND “lexicon” AND (“indonesia” OR “indonesian”)

The articles criteria for SLR are as follows:

1. Articles are published within the year 2020-2021.
2. The articles have the main topic of the sentiment analysis process, containing all or some of the following: the pre-processing stage, the modelling stage (knowledge), classification, and displaying classification accuracy.
3. The object of research is many domains, especially election of head of state/regional head, senate members, or legislative.

The excluded criterion of article for SLR:

1. Article that published before 2020.
2. Articles that the main subject is not sentiment analysis.
3. Data is not on Indonesian Language.

Based on the searching that excludes the criterion of a query and terms, we have found 30 articles.

c. Article Quality Assessment

The selected articles should be checked to meet the specified quality criteria. The questions to be used for the quality test are shown in Table 2.

Table 2: Quality Assessment Question

No.	Question			
1.	Does the study describe the purpose of sentiment analysis clearly?			
	Answer:	Yes(1)	Partly(0.5)	No(0)

2.	Does the study describe the pre-processing stage clearly?			
	Answer:	Yes(1)	Partly(0.5)	No(0)
3.	Does the study describe lexicon used clearly?			
	Answer:	Yes(1)	Partly(0.5)	No(0)
4.	Does the study describe the classification process using resulting model clearly?			
	Answer:	Yes(1)	Partly(0.5)	No(0)
5.	Does the study describe accuracy level clearly?			
	Answer:	Yes(1)	Partly(0.5)	No(0)

After undergoing the assessment of article quality, there are 10 articles that meet the purpose of writing this literature review.

RESULT AND DISCUSSION

a. Review of preprocessing

There are several studies that can be considered to reveal the best preprocessing of text. Research from (Tho et al., 2021)aims to compare SentiNetWord and VADER in extracting the polarity of the code-mixed sentences in Indonesian language and Javanese language from twitter. This study using several methods for pre-processing such as removing duplicates, translating to English, filter special characters, transform lower case and filter stop words were conducted on the sentences. Another research is (Sanjaya & Lhaksmana, 2020) using POS Tagging, Slangword replacement, emoticon and emoji detextion, case folding, and tokenizing. Similar research was conducted by (Prayoga et al., 2020), where preprocessing uses case folding, stemming, tokenizing, normalization, cleaning and santizing. This research is complemented by (Dikiyanti et al., 2021) using clear HTML and URL tags, stemming, and tokenizing. Research from (Firdaus et al., 2021) also uses tokenizing, case conversion, stemming, remove punctuation and remove stopword. Research (Aribowo & Khomsah, 2021) using remove hashtags, remove unicode strings, lowercase, tokenizing, remove number, convert slangwords, and remove stopwords

In conclusion, the preprocessing methods commonly used in lexicon-based sentiment analysis studies in Indonesian are:

1) Tokenization: the sentence breaks into a list of words and is space-delimited.
2) Case Conversion: Convert to lowercase letters so they are easy to process.
3) Stemming: change each word to root word.
4) Remove Punctuation : Eliminates punctuation in opinion.
5) Remove Stop Word : Eliminates words that do not contain sentiment elements.
6) Remove or replace emoji and emoticons into text that represents them
7) Normalization or slangword conversion to standard words.

b. Review The Labeling Process

In research (Tho et al., 2021) positive and negative words are determined from the lexicon model, then calculated using a simple mathematical formula in order to classify the polarity. By comparing with the manual labeling, the result showed that Senti Net Word performed better than VADER in negative sentiments but did not perform well in neutral and positive sentiments. Research (Hayaty et al., 2020) proposed method for labeling. The trick is to calculate the polarity of a text, the polarity score of the text will be checked. If the text is presented in the dictionary, then the score is added to get an overall polarity score. For example, if a text matches a word marked as positive in the dictionary, then the total polarity score of the text is increased. If the overall polarity score of a text is positive, then that text is classified as positive. Otherwise, it is classified as negative. In research (Prayoga et al., 2020) Lexicon based are divided into two data, namely positive word data and negative word data that has been labeled and grouped manually by humans. Research (Hernikawati, 2021) used Textblob by looking at the polarity. For polarity

the weighting value is between -1 to 1. The polarity value of -1 means negative sentiment and for polarity 1 value means positive sentiment.

The sentiment labeling process in these studies uses a manual calculation formula on average. Sentiment is calculated based on the comparison of the number of negative sentiment keywords with positive sentiment keywords in one sentence. This calculation does not pay attention to the semantic meaning of the sentence, because it is possible that these sentences actually have negative sentiment polarity but are recognized as sentences with positive or neutral sentiments.

c. Review of Lexicon

The review of the lexicon intends to find out the lexicon used in sentiment analysis studies in Indonesian. Some of these studies are from (Tho et al., 2021) whose research aims to compare two lexicon models which are SentiNetWord and VADER in extracting the polarity of the code-mixed sentences in Indonesian language and Javanese language. Both Lexicon are compared and VADER provides higher accuracy. Another sentiment lexicon research using 2 dictionaries for feature extraction using SentiWordNet and InSet, Indonesia Sentiment Lexicon (Sanjaya & Lhaksmana, 2020). Another research is from (Hernikawati, 2021) using Text Blob for sentiment lexicon. The TextBlob method can only process textual data in English, so in this study it was necessary to translate twitter opinion from Indonesian-language into English. Next research from (Musfiroh et al., 2021) and (Firdaus et al., 2021) also uses InSet Lexicon which produces a classification model in 3 classes, namely positive, negative, and neutral. InSet Lexicon has been tested quite well for sentiment analysis in Indonesian. InSet Lexicon (Indonesia Sentiment lexicon) consists of 3,609 positive words and 6,609 negative words that have a polarity score on each word with a weight between -5 to +5. This polarity score is used to classify the types of sentiment. Another research using The NRC Affect Intensity Lexicon is chosen in this research because it provides a wider range of emotions to analyze the news tweets, compared to using binary sentiment lexicon, i.e., positive and negative sentiments only. Furthermore, the lexicon also provides real-valued scores for rating the intensity of the 8 basic emotions (Suryadi, 2021). NRC also used by (Aribowo & Khomsah, 2021)for emotion detection.

Based on the results of the literature review above, the lexicon is used to extract sentiment features directly from the sentence to be tested. Each word in the opinion will be matched with a word on the lexicon. If the word is found, the sentiment will be searched on the lexicon. This process has the advantage of being simple. The downside is that a word can have multiple polarities, especially if it's been stemming. For example, the word "laughing" has a negative polarity. If the stemming process is carried out, it will become "laughter" or "laugh" with positive polarity.

d. Review of Accuracy

The results of the Tho's study explain that on the overall performance, Lexicon VADER provides better performance than SentiNetWord (Tho et al., 2021). Another research comparing accuracy using Indonesian lexicon sentiment with SentiWordNet. The result is the F-measure value of Indonesian lexicon sentiment (0.598) higher than SentiWordNet (0.413). Musfiroh's research tested the classification with cross-validation and confusion matrix with 80% training data and 20% test data. The results give an accuracy value of 79.2%, precision of 72.9%, recall of 62.8%, and f-measure of 67.4% (Musfiroh et al., 2021). Research from Firdaus on sentiment analysis of student feedback evaluation with InSet Lexicon was able to provide 90.9% accuracy in document level (Firdaus et al., 2021).

Based on observations on the accuracy of the sentiment analysis model using the lexicon, the accuracy can be quite high, reaching 90%. The high accuracy is proportional to the amount of preprocessing used, the lexicon used, and also the condition of the opinion dataset being assessed. So it is rather difficult to compare which model is the best, unless the model is tested on the same dataset.

CONCLUSION

Researches on lexicon-based Indonesian sentiment analysis in the last two years still use the old methods. The preprocessing techniques used are still using standard methods. The calculation of positive and negative sentiment scores still uses simple calculations, and the maximum accuracy of the classification results is 90%. The most widely used lexicon is NRC which is the result of translation from English and Inset which is a lexicon dictionary in Indonesian. Further research is how this lexicon method can also understand the semantic meaning of a sentence. How to enrich the Indonesian sentiment polarity

dictionary, and create other dictionaries that support the preprocessing stage, such as stop word dictionaries, slangword dictionaries and so on.

REFERENCES

Aribowo, A. S., & Khomsah, S. (2021). Implementation Of Text Mining For Emotion Detection Using The Lexicon Method (Case Study: Tweets About Covid-19). *Telematika*, *18*(1), 49. https://doi.org/10.31315/telematika.v18i1.4341

Dikiyanti, T. D., Rukmi, A. M., & Irawan, M. I. (2021). Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm. *Journal of Physics: Conference Series*, *1821*(1). https://doi.org/10.1088/1742-6596/1821/1/012054

Firdaus, R., Asror, I., & Herdiani, A. (2021). *Lexicon-Based Sentiment Analysis of Indonesian Language Student Feedback Evaluation*. *6*(April), 1–12. https://doi.org/10.34818/indojc.2021.6.1.408

Hayaty, M., Adi, S., & Hartanto, A. D. (2020). Lexicon-Based Indonesian Local Language Abusive Words Dictionary to Detect Hate Speech in Social Media. *Journal of Information Systems Engineering and Business Intelligence*, *6*(1), 9. https://doi.org/10.20473/jisebi.6.1.9-17

Hernikawati, D. (2021). *Kecenderungan Tanggapan Masyarakat Terhadap Vaksin Sinovac Berdasarkan Lexicon Based Sentiment Analysis The Trend of Public Response to Sinovac Vaccine Based on Lexicon Based Sentiment Analysis*. *23*(1), 21–31.

Khomsah, S., & Aribowo, A. S. (2020). Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia. *Rekayasa Sistem Dan Teknologi Informasi, RESTI*, *4*(10), 648–654. https://doi.org/https://doi.org/10.29207/resti.v4i4.2035

Musfiroh, D., Khaira, U., Eko, P., Utomo, P., & Suratno, T. (2021). Sentiment Analysis of Online Lectures in Indonesia from Twitter Dataset Using InSet Lexicon Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(April), 24–33.

Prayoga, N. R., Tresna Maulana Fahrudin, Made Kamisutara, Rahagiyanto, A., Primananda Alfath, T., Latipah, Winardi, S., & Susilo, K. E. (2020). Unsupervised Twitter Sentiment Analysis on The Revision of Indonesian Code Law and the Anti-Corruption Law using Combination Method of Opinion Word and Agglomerative Hierarchical Clustering. *EMITTER International Journal of Engineering Technology*, *8*(1), 200–220. https://doi.org/10.24003/emitter.v8i1.477

Sanjaya, G., & Lhaksmana, K. M. (2020). Analisis Sentimen Komentar YouTube tentang Terpilihnya Menteri Kabinet Indonesia Maju Menggunakan Lexicon Based. *E-Proceeding of Engineering*, *7*(3), 9698–9710.

Sasmito Aribowo, A. (2018). Analisis Sentimen Publik pada Program Kesehatan Masyarakat menggunakan Twitter Opinion Mining. *Seminar Nasional Informatika Medis (SNIMed)*, *0*(0), 17–23. https://journal.uii.ac.id/snimed/article/view/11877

Suryadi, D. (2021). Does it make you sad? A lexicon-based sentiment analysis on COVID-19 news tweets. *IOP Conference Series: Materials Science and Engineering*, *1077*(1), 012042. https://doi.org/10.1088/1757-899x/1077/1/012042

Tho, C., Heryadi, Y., Lukas, L., & Wibowo, A. (2021). Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach. *Journal of Physics: Conference Series*, *1869*(1), 8–14. https://doi.org/10.1088/1742-6596/1869/1/012084