International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 32 No. 2 May 2022, pp. 277-285

Sentiment Analysis Research in Indonesian Language Reviewing From the Characteristics of Comments

M. Isnin Faried[1], Lely Priska D. Tampubolon[*2], Dwi Atmodjo [3]

[1,2,3]Faculty of Information Technology
Perbanas Institute
Jakarta, Indonesia

Abstract – This study focuses on identifying the method of sentiment analysis using the data sources from different social media. The comments are classified based on the themes: politics, business product reviews, events, etc. The study also focuses on the form of the language and the data types: text only and texts with emoticons. The literature review is conducted on several previous studies with characteristics: Indonesian language, sarcasm, the method used, and the development of certain features in the sentiment analysis method. There are three conclusions. First, there are polite comments and impolite/ sarcastic comments. Comments posted on government channels are more polite than those on social media channels. All the comments have the same polarity. The comments use words only or a combination of words and emoticons. The analysis becomes complex because the comments use slang words. Second, the support vector machine (SVM) method is widely used. The use of libraries in doing sentiment analysis in Indonesian is helpful, but only a few are suitable for general purposes. Finally, the feature development has many variants which can be customized based on the needs, and SentiWordNet is the most popular supporting application.

Keywords – ***Comments, Dataset, Feature Development, Sentiment Analysis***

I. INTRODUCTION

Social media has been widely used as a channel for the community to express their opinions, especially to express what they feel about an event or a fact that they have seen and experienced. Social media has developed fairly rapidly since the access to the internet has become easier [1]. On social media, public opinion can be referred to any posted comments which can have an influence on the readers. This influence can be observed, particularly for comments which are related to certain products or to public figures which can be seen during presidential elections and local government elections.

Seeing the magnitude of the influence resulting from the comments, it is necessary to do an automatic analysis of comments, specifically to identify whether or not the comments are positive or negative [1]. To analyze comments, it is common to use the sentiment analysis which is a method to study the computation of public comments, attitudes, and emotions towards an entity, in the forms of individuals, certain topics, and events that occur around the comments [2], [3].

The objects of this study are comments taken from social media. With the large volume and variance of the data in the forms of certain text symbol, the objects of this analysis can be considered as big data. The big data analysis process is applied as a solution when the manual process of data and the old technology cannot be used to process very large data [4].

Many studies on sentiment analysis have been carried out with the aim of identifying the appropriate method used for various dataset sources. The data source variants in question are data sources that come from different social media, and the data sources of comments can also be classified based on the themes, for example comments about politics, about business product reviews,

Corresponding Author: Lely Priska D. Tampubolon

about events, etc. Another source that is no less important is the dataset in the form of the language used in comments. Furthermore, there are also studies that focus on the types of comments, such as the use of text symbols or images showing emotion known as emoticon. Based on the type of research, a literature review of the research is conducted in order to analyze the characteristics of the existing data sources and the methods used based on these characteristics.

This study reviews several previous studies that focus on the characteristics of the dataset: Indonesian language, sarcasm, the method used, and the development of certain features applied to the sentiment analysis method on their studies. The previous studies were obtained from several research publications between 2013 and 2021. This article is divided into the introduction, the literature review, the research method, the results and discussions which contains the answers to the research questions (RQ), and the conclusion.

II. RESEARCH METHODS

This study adopted some of the systematic literature review methods proposed by B. Kitchenham and S. Charters [5] who define Systematic Literature Review (SLR) as a method used in identifying, evaluating, and interpreting all research results which can be used as references to answer certain research questions [6].

2.1 Research Questions (RQ)

Research questions are important in the literature review because they can make the review more focused and have a clear direction. RQ is made based on PICOC criteria: Population (P), Intervention (I), Comparison (C), Outcomes (O), and Context (C). Population refers to the target group; intervention states the focus of observation in research; comparison is used when there are stages of doing comparisons on the intervention; outcomes indicate the impact caused by the intervention; and context defines the scope of the study. The research questions of this study are shown in Table 1 and Table 2.

Table 1. Research Question Criteria

Criteria (PICOC)	Items
Population	Sentiment analysis, opinion mining
Intervention	Datasets, methods
Comparison	n/a
Outcomes	Indonesian language dataset characteristics, feature development
Context	Social media

Table 2 . Research Questions in Literature Review

Research Question	Motivation
What are the characteristics of Indonesian language datasets?	Identify the characteristics of comments in Indonesian language
What method is used?	Identify the method used by the researchers for the Indonesian language dataset
What feature development is done?	Identify developments carried out in order to conduct the sentiment analysis with Indonesian language datasets

2.2 The Search Strategy for the Articles

The analyzed articles were taken from various sources which are selected based on several reputable indexing databases as one of the criteria. Other criteria specify that article sources are from reputable universities, and, most importantly, they must study Indonesian language datasets. The articles are published between 2013 and 2021.

2.3 Literature Selection

To select the literature, there are two criteria: inclusion and exclusion, and the results are shown in Table 3. The properties of the literature review are displayed in Table 4. The list of articles is shown in Table 5.

Table 3. Literature Selection Criteria

Criteria	Selection Items
Inclusion	Research using datasets from social media
	Research using Indonesian language dataset
Exclusion	Research using the lexicon and not using the lexicon
	Feature development of sentiment analysis method

Table 4. Literature Review Properties

Property	Research Question
Data source	[RQ1] What are the characteristics of the Indonesian dataset?
Method	[RQ2] What method is used?
Development Method	[RQ3] What feature development is conducted?

Table 5. List of Articles

Title	Researcher	Article Type
A survey of sentiment analysis using SentiWordNet on bahasa Indonesia [7]	S. Christina, and D. Ronaldo	Journal
Dataset Indonesia untuk analisis sentimen (Indonesian dataset for sentiment analysis) [8]	R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka	Journal
Indonesian social media sentiment analysis with sarcasm detection [9]	E. Lunando, and A. Purwarianti	Journal
Analisis sentimen pada Twitter mahasiswa menggunakan metode backpropagation (Sentiment analysis on student Twitter using backpropagation method) [10]	R. Habibi, D. B. Setyohadi, and E. Ernawati	Journal

Sentiment analysis Twitter bahasa Indonesia berbasis Word2vec menggunakan Deep Convolutional Neural Network (Sentiment analysis of Indonesian Twitter based on Word2vec using Deep Convolutional Neural Network) [11]	H. Juwiantho, E. I. Setiawan, J. Santoso, and M. H. Purnomo	Journal
Penerapan analisis sentimen pada Twitter berbahasa Indonesia sebagai pemberi rating (The application of sentiment analysis on Twitter in Indonesian language as the rater) [12]	N. Monarizqa, L. E. Nugroho, and B. S. Hantono	Journal
Sentiment analysis berbasis big data (Sentiment analysis based on big data) [4]	P. Nomleni, M. Hariadi, and I. K. E. Purnama	Proceedings
Sentiment analysis terhadap Tweet bernada sarkasme berbahasa Indonesia (Sentiment analysis against sarcasm Tweets in Indonesian) [13]	L. Septiani, and Y. Sibaroni	Journal

2.4 Data Extraction

The data were extracted from the selected articles as listed in Table 5. This data extraction was conducted for the purpose of answering the research questions (RQ) of this study. The obtained data are mapped onto the property in order to answer the existing RQ. The stages from the formulation of research questions to data extraction are adopted partially from the existing literature review stages.

2.5 Literature Analysis and Discussion

Sentiment analysis (SA) is a process of analyzing the text-based information which is abundant on the internet. The perceived results of the analysis are in the forms of information that can be used as knowledge for organizations and individuals. The results can be beneficial to organizations which can make certain evaluation in order to make improvements; similarly, individuals can use the results to do various things, including making decisions [1]. The basic idea of this SA is a conclusion to the comments which are categorized into positive, neutral, and negative comments. The analyzed comments are a collection of texts that have been posted on various digital platforms (or the sources) on the internet [14].

The application of sentiment analysis always uses data sources that become the objects of analysis; it constitutes the methods used in the analysis and the development of methods that are applied to solve certain problems. This method development in question is to conducted to select the features of the SA method. The selection of features can provide a better level of accuracy in the sentiment analysis process [15].

As mentioned earlier, this study conducted a literature review of sentiment analysis articles which used Indonesian language datasets as the objects of discussion, the methods which are used, and the development of features (feature selection). The results of the analysis are shown in Table 6.

III. Results And Discussions

3.1 Characteristics of Comments in Indonesian Language (RQ1)

Research [7] with the Indonesian language dataset as its object concludes that comments with a polite tone create the sentiment polarity with multiple meanings; consequently, the accurate results of the analysis cannot be found immediately. These

polite comments are often found on channels on government websites; on the other hand, comments which tend to be negative can easily be found on social media channels.

The results of Research [12] show that there is a possibility of having duplicate comments and spam. Therefore, it is necessary to do the data filtering at the outset. Furthermore, it also concludes that there are certain characteristics of Indonesian comments in that they use a combination of letters and numbers to form a word and an over-repetition of certain letters in a word. Certain numbers are used to replace certain letters, for example in the word "*saya*" (meaning 'I') which is written as "s4ya" in some comments. Another example shown here is the interjection word "*semangat*" (meaning 'keep your spirit up') which is written as "*semangaaattt*" with the over-repetition of the letter {a} and {t}. In addition to analyzing the texts in the comments, this study also examines the emoticons which accompany the texts in the comments.

Research [4] has similar results to Research [12] as they also find the combination of letters and numbers to form a word and an over-repetition of certain letters in a word. However, Research [4] find instances of slank words instead of proper and formal words used in the comments. For examples, some comments contain the word "*bingit*" (meaning 'very much') which is a slang word derived from the less formal Indonesian word "*banget*", which actually has the more formal words such as "*sangat*" or "*sekali*" (meaning 'very much').

Research [9] and [15] yield nearly similar results as both research focus on comments with sarcastic tone. In general, many sarcastic sentences are used in the comments which review certain products and give opinions about issues related to the government and politics. The results are also in line with those of Research [7] in which the comments are posted on social media channels.

Table 6. Article Analysis

Research Title and Author	Dataset	Method
A survey of sentiment analysis using SentiWordNet on Bahasa Indonesia, A: S. Christina, and D. Ronaldo.[7]	Indonesian (533)	Lexical, SentiWordNet
Notes: SentiWordNet has sentiment scores: Pos(s), Neg(s) and Obj(s). SentiWordNet does not support Indonesian subtitles, so a translation process is needed from Indonesian to English. Lexical selection, unsupervised method, to avoid the problem of limited subjects as well as the problem of manual data labeling which takes a long time. Comments used in Indonesian are more polite, causing more than one polarity of sentiment and creating ambiguity. Comments through government channels are more polite compared to Twitter and Facebook channels which tend to show negative sentiments. The accuracy of the results is higher if the sentiment polarity score is given on the lexical database.		
Sentiment analysis Twitter bahasa Indonesia berbasis Word2vec menggunakan Deep Convolutional Neural Network (Sentiment analysis of Indonesian Twitter based on Word2vec using Deep Convolutional Neural Network) A: H. Juwiantho, E. I. Setiawan, J. Santoso, and M. H. Purnomo. [11]	Indonesian (999), Text and image data	Deep Convolutional Neural Network (Deep Learning), Word2Vec Model.

	Notes: The application of Word2Vec-based deep convolutional neural network is better for recognizing sentiment in texts by taking into account word sequences in sentences. This advantage is an improvement from the classical method which ignores word order. Indonesian Word2Vec is used to initialize words into their vector forms so that in the process, no training and manual feature searches are required. Word2Vec creates vectors for words that have similarities to make analysis easier. The research focuses on the implementation of the method and does not pay attention to the dataset.	
Penerapan analisis sentimen pada Twitter berbahasa Indonesia sebagai pemberi rating (Application of sentiment analysis on Twitter in Indonesian language as the rater), A: N. Monarizqa, L. E. Nugroho, and B. S. Hantono.[12]	Indonesian, texts and emoticons (175,000)	Support Vector Machine (SVM), Lucene Library
	Notes: Analysis of emoticons using a conversion table. There have been findings of duplicate comments and spam. Existing comments are characterized by consecutive numbers and letters.	
Analisis sentimen pada Twitter mahasiswa menggunakan metode backpropagation (Sentiment analysis on student Twitter using the backpropagation method) A: R. Habibi, D. B. Setyohadi, and E. Ernawati.[10]	Indonesian (6000)	Backpropagation as classifier, WEKA multilayer perceptron
	Notes: The initial results in the classification stage were deficient because the manual labeling resulted in multiple interpretations.The sentiment dictionary is lacking in vocabulary and emoticons.	
Dataset Indonesia untuk analisis sentimen (Indonesian dataset for sentiment analysis) A: R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka.[8]	Indonesian (454,559)	Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD)
	Notes: The existing Indonesian datasets tends to be more specific and in line with the topics of the research; thus, they are not universal. The 3 algorithms used can show that the created dataset has higher values than the comparative data.	

Sentiment analysis berbasis big data (Sentiment analysis based on big data) A: P. Nomleni, M. Hariadi, and I. K. E. Purnama.[4]	Indonesian	Support Vector Machine (SVM)
Notes: Comments in Indonesian on social media tend to use uncommon words, for example replacement of vowels with numbers, repetition of vowels, and the use of slang words. The data acquisition stage adds the Indonesian language detection process to obtain more data. Utilizing big data and processed using Hadoop Distributed File System (HDFS). Higher accuracy can be achieved if there were more data.		
Indonesian social media sentiment analysis with sarcasm detection A: E. Lunando, and A. Purwarianti. [9]	Indonesian, Sarcasm	Naïve Bayes, Maximum Entropy, Support Vector Machine (SVM), Unigram, Sentiwordnet
Notes: The sarcastic comments are more often conveyed on the topic of products, government, and politics. The addition of negativity and interjection features provides an effective analytical work for sarcastic words.		
Sentiment analysis terhadap Tweet bernada sarkasme berbahasa Indonesia (Sentiment analysis against sarcastic Tweets in Indonesian) A: L. Septiani, and Y. Sibaroni.[15]	Indonesian, Sarcasm	Supervised Learning, Naïve Bayes, Support Vector Machine (SVM)
Notes: Higher classification accuracy is achieved by using Naïve Bayes. The interjection feature has a positive effect in terms of increasing classification accuracy. It is better to apply the interjection and unigram features in analyzing comments with sarcastic tones.		

3.2 The Methods Used in the Indonesian Language Dataset (RQ2)

There are a number of methods used in sentiment analysis, but from the literature analysis, it is found that there are several methods which are commonly used in several studies related to Indonesian-language objects. The common methods used in the literature are Deep Learning by using Deep Convolutional Neural Network, Support Vector Machine (SVM), Backpropagation, K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), Naïve Bayes, and Maximum Entropy.

Besides the methods, several studies also provide additional information in the form of libraries used in research, especially for Indonesian datasets. The Lucene library was used by Research [12], while Research [8] tried to create its own library as an addition to the existing library in the study, namely SemEval-2018.

3.3 Development of Features in Sentiment Analysis Research with Indonesian Language Datasets (RQ3)

Research [7] uses a lexical-based research pattern, unsupervised method, and SentiWordNet tools. Lexical and unsupervised methods were chosen in this study to avoid the problem of having limited number of subjects and of doing data labelling manually which takes a long time. SentiWordNet is used because it has a sentiment score. The final results of this study show that this method yields higher accuracy. This is good because it includes the sentiment polarity scores in the lexical database.

Research [11] uses the Deep Convolutional Neural Network pattern based on Word2Vec. Indonesian Word2Vec is used to initialize words into their vector forms so that in the process no training and manual feature searches are required. Word2Vec creates vectors for words that have similarities to make analysis easier. The results of this research offer the benefit of making improvements from the classic method that ignores the word order in a language.

Research [10] uses a pattern by utilizing the backpropagation method as the classifier and WEKA multilayer perceptron. The utilization of WEKA can reduce deficiencies during the classification stage resulting from manual labeling which creates multiple interpretations.

Research [4] uses the support vector machine (SVM) method and the Hadoop Distributed File System (HDFS) for big data processing. This research also uses the language detection process at the acquisition stage, and the results show a higher accuracy because there are more training data.

Research [15] uses a method that is nearly the same as that of Research [9] because they examine sarcastic comments. The research models used are unigram, SentiWordNet which utilizes Support Vector Machine (SVM) and Naïve Bayes. This research also uses additional features of negativity and interjection. The results of this study indicate that the Naïve Bayes method has higher accuracy, and the interjection feature has a positive effect in terms of increasing the accuracy of the classification. Moreover, it is better to use the interjection and unigram features in analyzing texts with sarcastic tones.

IV. Conclusion

There are three parts (3 RQ) in the conclusion of this literature analysis which is conducted in order to answer the 3 research questions (RQ).

First, with regard to the Indonesian language dataset, it can be concluded that comments in Indonesian have two features: polite comments and impolite (or sarcastic) comments. Comments posted on government channels are more likely to be polite than those posted on social media channels. The polite comments and sarcastic comments have the same polarity and require certain ways to produce an accurate analysis. The comments are conveyed by using words only or a combination of words and emoticons. It takes more effort to conduct the analysis which becomes more complex because the comments are written in less formal words (or slangs).

Second, it can be concluded that the support vector machine (SVM) method is widely used. The use of libraries in doing sentiment analysis in Indonesian language is helpful because there are so many libraries; however, many of them are topic-specific in order to be used in specific research, so only a few are suitable for general purposes.

Finally, it can also be concluded that the feature development has many variants which can be customized based on the needs. In addition, SentiWordNet is the most popular supporting application that are widely used in many studies.

V. Acknowledgment

The authors would like to thank the Ministry of Technology and Higher Education which has supported this research with research funding. Finally, we would also like to thank Perbanas Institute, especially the Research and Community Service that always supports all the research activities.

Reference

[1] U. Hemamalini and S. Perumal, “Literature review on sentiment analysis,” *Int. J. Sci. Technol. Res.*, vol. 9, no. 4, pp. 2009–2013, 2020.

[2] C. Bhadane, H. Dalal, and H. Doshi, “Sentiment Analysis: Measuring Opinions,” in *Procedia Computer Science*, 2015, vol. 45, pp. 808–814.

[3] W. Medhat, A. Hassan, and H. Korashy, "Sentiment Analysis Algorithms and Applications: A Survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[4] P. Nomleni, M. Hariadi, and I. K. E. Purnama, "Sentiment Analysis Berbasis Big Data," in *Seminar Nasional Rekayasa Teknologi Industri dan Informasi*, 2014, vol. 9, pp. 142–149.

[5] B. Kitchenham and S. Charters, "Guidelines for Performing Systematic Literature Reviews in Software Engineering," 2007.

[6] R. S. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," *J. Softw. Eng.*, vol. 1, no. 1, pp. 1–16, 2011.

[7] S. Christina and D. Ronaldo, "A Survey of Sentiment Analysis Using Sentiwordnet on Bahasa Indonesia," *J. Teknol. Inf.*, vol. 12, no. 2, pp. 169–174, 2018.

[8] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 334, 2019.

[9] E. Lunando and A. Purwarianti, "Indonesian Social Media Sentiment Analysis with Sarcasm Detection," *2013 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2013*, pp. 195–198, 2013.

[10] R. Habibi, D. B. Setyohadi, and E. Wati, "Analisis Sentimen Pada Twitter Mahasiswa Menggunakan Metode Backpropagation," *J. Inform.*, vol. 12, no. 1, pp. 103–109, 2016.

[11] H. Juwiantho *et al.*, "Sentiment Analysis Twitter Bahasa Indonesia Berbasis WORD2VEC Menggunakan Deep Convolutional Neural Network," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 1, pp. 181–188, 2020.

[12] N. Monarizqa, L. E. Nugroho, and B. S. Hantono, "Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating," *J. Penelit. Tek. Elektro dan Teknol. Inf.*, vol. 1, no. 3, pp. 151–155, 2014.

[13] L. Septiani and Y. Sibaroni, "Sentiment Analysis Terhadap Tweet Bernada Sarkasme Berbahasa Indonesia," *J. Linguist. Komputasional*, vol. 2, no. 2, pp. 62–67, 2019.

[14] R. Kumar, D. Sarddar, I. Sarkar, R. Bose, and S. Roy, "A Literature Survey On Sentiment Analysis Techniques Involving Social Media And Online Platforms," *Int. J. Sci. Technol. Res.*, vol. 9, no. 05, p. 5, 2020.

[15] F. Septianingrum and A. S. Y. Irawan, "Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 799, 2021.