Lexicon-Based Sentiment Analysis Using Inset Dictionary: A Systematic Literature Review

Asy Syifaur Roisah Rufaida [1], Adhistya Erna Permanasari [2], Noor Akhmad Setiawan [3]

{ asysyifaurroisahrufaida@mail.ugm.ac.id [1], adhistya@ugm.ac.id [2], noorwewe@ugm.ac.id [3]}

Department of Electrical Engineering and Information Technology Universitas Gadjah Mada, Yogyakarta, Indonesia[1], Department of Electrical Engineering and Information Technology Universitas Gadjah Mada, Yogyakarta, Indonesia[2], Department of Electrical Engineering and Information Technology Universitas Gadjah Mada, Yogyakarta, Indonesia[3]

Abstract. Sentiment analysis is one of the most exciting topics to research in text mining. Lexicon-based sentiment analysis produces good results across a wide range of conversational topics, is easily improved using various sources of information, and does not require additional training. However, only a few review papers have addressed sentiment analysis with specific lexicon dictionaries. As a result, neither data sources nor pre-processing techniques specific to a lexicon dictionary are discussed further. A Systematic Literature Review was conducted to provide a comprehensive and structured lexicon-based sentiment analysis using the Inset dictionary. The literature review resulted in selecting seventeen papers for a detailed study. The findings show that in the last five years, most sentiment analysis research using Inset focused on the health domain. Twitter provides the majority of the data for sentiment analysis. Stopword removal, tokenization, case folding, and stemming are common pre-processing techniques. This study also contained some additional observations from completed research.

Keywords: Sentiment Analysis, Lexicon-Based Method, Indonesia Sentiment Lexicon.

1 Introduction

Sentiment analysis is one of the most intriguing issues to investigate in text mining [1]. Sentiment analysis takes features from structured or unstructured textual data and evaluates them to derive opinions, emotions, and feelings [2]. The fundamental advantage of sentiment analysis is that it categorizes opinions as positive, negative, or neutral. The correct tools and methodologies are required to assess these viewpoints. A suitable sentiment analysis method will produce sentiment analysis with high accuracy under actual conditions.

There are two methods for extracting sentiment from the opinion: lexicon-based sentiment analysis and machine-learning-based sentiment analysis [1]. Lexicon-based sentiment analysis is the conventional approach [3]. This approach scans each sentence for terms that convey a positive or negative emotion. This method yields good performance across a range of conversational subjects, which can be easily improved using various sources of information and does not call

ICAE 2022, October 05, Batam, Indonesia

DOI 10.4108/eai.5-10-2022.2327474

for additional training. There are two primary techniques for creating sentiment lexicons: dictionary-based [4] and corpus-based [5], [6]. The dictionary-based method frequently succeeds in the public domain. Meanwhile, corpus-based lexicons can be tailored to specific domains [7].

Sentiment dictionaries can act as a word-level basis; the lexicon evaluates the sentiment of unlabelled texts or sentences to gather specific information like the polarity and strength of the words contained therein. Every word in the sentiment lexicon has a consistent polarity. However, if the word is used in a different domain, it has a different polarity and, as a result, a different power weight [8]. This indicates that a sentence's sentiment is determined using specific calculations based on knowledge from a language dictionary [1]. Various lexicon dictionaries include Liu, SentiWordNet, and LIWC (Linguistic Inquiry and Word Count). These dictionaries are English lexicon dictionaries. In addition, digital Indonesian dictionaries are being developed [1]. Several Indonesian dictionaries are available, including Inset (Indonesia Sentiment Lexicon) [9], Masdevid, and Barasa.

Although there are several review papers on sentiment analysis, the vast majority discuss sentiment analysis in general. Fauziah et al. [1] examined the last two years of lexicon-based sentiment analysis research in the Indonesian Language. The study focuses on the understanding of pre-processing used in recent lexicon-based sentiment analysis studies, the lexicon used in these studies, and classification accuracy. Significantly few review papers have addressed sentiment analysis using specific lexicon dictionaries. As a result, neither data sources nor pre-processing techniques appropriate for a specific lexicon dictionary are discussed further.

This paper aims to identify and analyze recent research on the use of the Inset dictionary in lexicon-based sentiment analysis. The Inset dictionary is one of the most extensively used Indonesian lexicon dictionary in sentiment analysis. This study focuses on understanding the domain discussed in the last five years, the data sources used in the studies, and pre-processing used in sentiment analysis using the Inset dictionary. In this literature review, we will not go into greater detail about the lexicon method and its accuracy. A Systematic Literature Review (SLR) approach was used to identify research, methodology, theoretical, and empirical gaps in sentiment analysis that can be addressed with further research.

2 Methodology

The method of a systematic literature review is used in this investigation. This method is used to investigate a problem in depth to identify gaps in earlier research. This methodology locates and evaluates all relevant research to answer the research question. Figure 1 depicts the steps of this study. The three primary steps, namely the planning stage, the conducting stage, and the reporting stage, are described in the following sections.

2.1 Planning

Research questions are developed as this research's main limitation or focus throughout the planning stage. A research question for this study is as follows:

RQ1*:* What are the most popular domains discussed in sentiment analysis studies using the Inset dictionary?

RQ2*:* What data sources are used in sentiment analysis studies using the Inset dictionary?

RQ3*:* What are the most common pre-processing techniques used in sentiment analysis studies using the Inset dictionary?

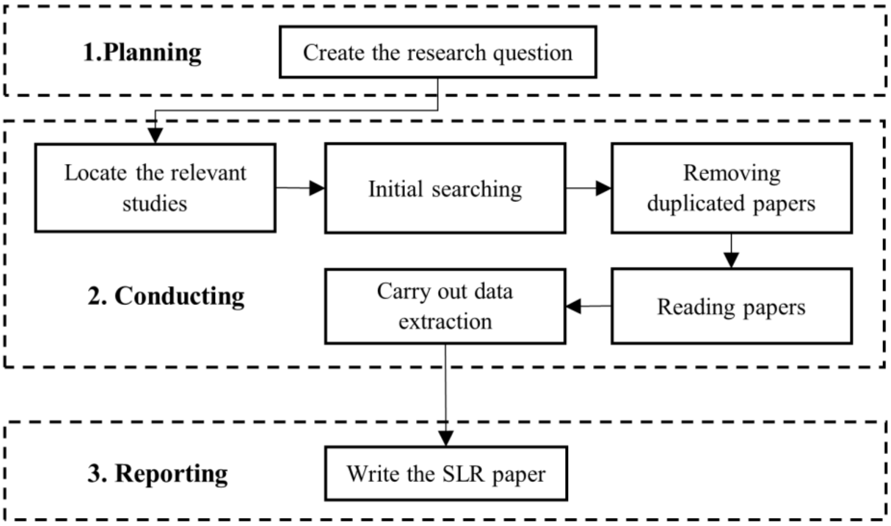
1.Planning
Create the research question
Locate the relevant studies
Initial searching
Removing duplicated papers
2. Conducting
Carry out data extraction
Reading papers
3. Reporting
Write the SLR paper

Fig. 1. Stages of Research Methodology

2.2 Conducting

A search for relevant literature began during the conducting stage. It started with an initial search on Google Scholar and the IEEE (Institute of Electrical and Electronics Engineers) digital library. The keywords "Inset (Indonesia Sentiment)" were utilized. After finding multiple articles with these keywords, the author reviewed and deleted articles with identical content. After eliminating duplicate papers, the results are assessed by reading the complete manuscript. Articles that did not match the criteria for inclusion (Table 1) were also removed from the list of publications. The extraction of data is the final stage of the conducting stage. The data from the selected relevant papers were used to address the research question in this study.

Table 1. Inclusion and Exclusion Criteria

Type	Inclusion	Exclusion
Document Type	All relevant articles in the library	Book OR Literature reviews OR General data
Topics	The central theme of the articles is the sentiment analysis process, which includes all or some of the following stages: pre-processing, modeling (knowledge), and classification.	Sentiment analysis studies that do not use an Inset dictionary

Type	Inclusion	Exclusion
Datasets	Studies using Indonesian Language datasets	
Year of publication	2018-2022	

2.3 Reporting

The data is gathered and documented in the form of a Systematic Literature Review paper after it has been extracted.

3 Results and Discussion

This study was carried out in July 2022 with the help of Google Scholar and the IEEE library. The initial search returned 62 publications using the keyword "Inset (Indonesia Sentiment)." After removing duplicated papers and reviewing them more thoroughly, the number of publications decreased to seventeen due to the applied exclusion criteria. The exclusion criteria limited the number of published lists of duplicate content, sentiment analysis that was not lexicon-based, non-Indonesian datasets, and themes that were not connected to sentiment analysis.

Figure 2 depicts the number of publications found during the initial search and the final results following a more thorough article selection. The initial search returned 62 publications with the following details: publications in 2018 obtained two papers, in 2019 obtained five papers, in 2020 obtained thirteen papers, in 2021 obtained thirty papers, and in 2022 obtained twelve papers. After that, a more in-depth selection was carried out, and the final number of papers to be reviewed was seventeen. The following are the details of the final results of the paper selection: for publication in 2018, two papers were obtained; in 2019, one paper was obtained; in 2020, there were five papers; in 2021, six papers were obtained; and in 2022, three papers were obtained.

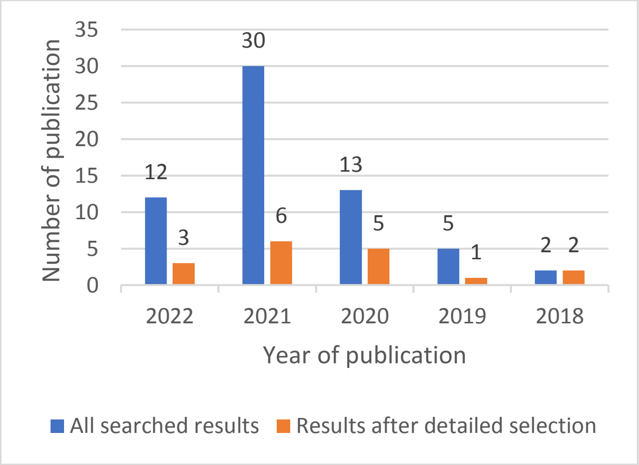

Fig. 2. Number of publications on sentiment analysis using Inset lexicon

Data extraction was then carried out to answer the research question based on the outcomes of this careful selection. The results of the data analysis in answer to the research question are listed below.

3.1 RQ1: What are the most popular domains discussed in sentiment analysis studies using the Inset dictionary?

According to the final selection results, the health domain is the most popular sentiment analysis. As shown in Fig. 3, Health domain papers accounted for five papers of the final selection [10]–[14]. Then followed by four papers in political domains [13], [15]–[17], two papers about App [7], [8], and three papers in general domain [18]–[20], as well as one each on another domain: education [21], regulation [22], disaster [23], and tourism [24]. The general domain is public sentiment without a specific topic taken within a particular time.

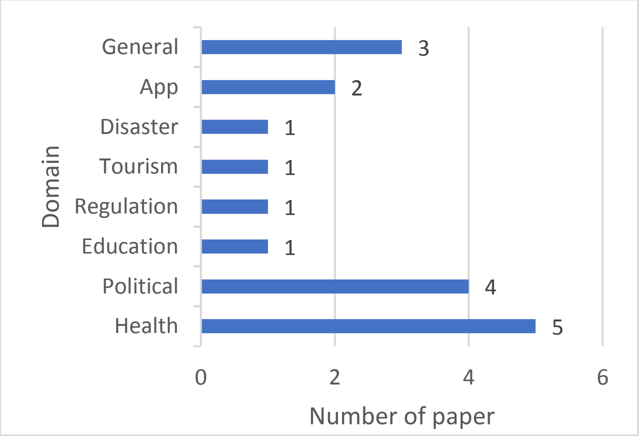

Fig. 3. Domains discussed in sentiment analysis studies using the Inset dictionary

Four publications in the health sector described the COVID-19 pandemic, but only one item discussed BPJS Kesehatan. BPJS Kesehatan (Badan Penyelenggara Jaminan Sosial Kesehatan, or Health Social Security Agency) is an Indonesian social security agency dedicated to providing universal health care to its citizens [25]. As seen in [12], Indonesians' sentiments toward BPJS Kesehatan were examined. This study employed topic modeling to determine which topics are most frequently discussed on Twitter. Whereas for COVID-19 topics, Jayapermana et al. [10] examined the sentiments of Indonesian netizens toward the COVID-19 vaccine. Another study [11] conducted public opinion clustering related to quarantine policy during the COVID-19 pandemic. The government can use these various viewpoints to make policy decisions. Saputra et al. [13] employed a dataset with two distinct topics: politics and health. Political topics discussed were the 2019 Presidential Election campaign and the 2018 West Java Governor Election campaign. Meanwhile, the health topics discussed were the lockdown and PSBB during COVID-19. Lockdown is a method of preventing the spread of a specific virus or disease by closing off all exit and entry points. PSBB (Pembatasan Sosial Berskala Besar) or Large-Scale Social Restrictions restricts certain activities of residents in areas suspected of being infected with COVID-19 to prevent the virus from spreading further. Mustofa and Prasetiyo [14] discussed the new normal era during the COVID-19 pandemic. The new normal policy entails resuming limited economic, social, and public activities while adhering to health standards that did not exist before the pandemic.

The author believes that the health domain, particularly about COVID-19, is dominant since it is the most appropriate circumstance for people's life nowadays. This is related to sentiment, primarily based on public opinion; thus, the health topic should be investigated further in the future.

3.2 RQ2: What data sources are used in sentiment analysis studies using the Inset dictionary?

As we can see in Fig. 4, the majority of the data for the sentiment analysis came from social media, especially Twitter. Twelve paper datasets were from Twitter [10], [11], [23], [24], [12]–[14], [16]–[19], [22], and one paper datasets each was from Instagram comments [15], student feedback [21], research answers [20], google play site [7] and play store [8].

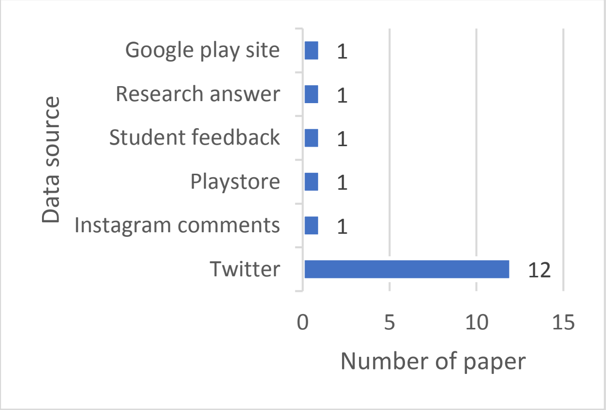

Fig. 4. Data sources used in sentiment analysis studies using the Inset dictionary

Twitter data crawling is done in several ways. According to [11], [14], [19], [23], they used the Twitter API that provides unique and advanced programmatic access to Twitter. Other studies used the python library for data crawlings, such as Tweepy, Twint, and GetOldTweets. As seen in [12], [13], [17], [18], Tweepy is a python library that requires a token to access the Twitter API. Tokens are obtained after applying to the Twitter app developer. Other studies [10], [24] used Twint, a tool for scrapping from Twitter applications that are specially set up using the Python programming language. We can use and run this library without having to use the API from Twitter itself. Suryadikara et al. [16] used the GetOldTweets library, which allows us to download historical Twitter data for a specific query within a specific date range. Furthermore, data extraction tools such as Octoparse [29] can be used. Octoparse is a cutting-edge visual web data extraction tool. The Twitter URL for the dataset is obtained and entered into Octoparse. After entering the URL, configure the data pagination process, which includes the ability to extract data from different pages.

Twitter is an excellent place to find sentiment datasets because it allows individuals to post about various current events and subjects. The tweet's content is similarly unrestricted; it can include ideas, critiques, or personal thoughts. Inset was also created using words from Twitter, so Twitter is most likely the appropriate data source for sentiment analysis using the Inset dictionary.

3.3 What are the most common pre-processing techniques used in sentiment analysis studies using the Inset dictionary?

The most frequently used pre-processing methods in sentiment analysis include:

a. *Stopword removal:* removes common irrelevant words in the sentiment classification process, such as prefixes, times, and conjunctions (example: "at, from, when, which, to", and so on.).
b. *Tokenization:* break sentences into single words.
c. *Case folding:* converts all letters into lowercase.
d. *Stemming:* converting words with affixes or compound words into their basic forms.
e. *Normalization:* correct words that are incorrectly spelled.
f. *Remove punctuation:* remove punctuation marks such as commas, periods, quotation marks, exclamation points, etc.
g. *Drop duplicate:* remove duplicate data.

Figure 5 depicts the number of articles processed using the pre-processing method. The stopword removal method was used in fifteen papers [17], [26]–[31], [18]–[25], tokenization was used in twelve papers [8], [10], [23], [24], [11]–[13], [15], [17], [20]–[22], case folding was used in thirteen papers [10], [11], [22]–[24], [12]–[14], [17]–[21], stemming was used in ten papers [8], [10], [12], [15], [18]–[22], [24], normalization was used in nine papers [10], [13], [15]–[19], [23], [24], and removing punctuation was used in eight papers [10]–[12], [14], [18], [20]–[22]. Drop duplicate was used in three papers [8], [16], [24].

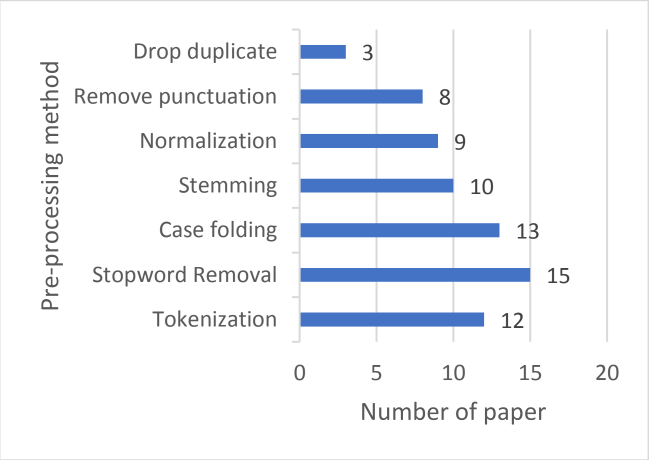

Fig. 5. Number of articles processed using the pre-processing method

Data pre-processing was required because the collected data was still unstructured, with the contents of each data still written in a non-standard language [17]. Several studies [11], [13], [16]–[18] used stopword lists to aid in the stopword removal process. Tala's Bahasa Indonesia stopword is used in those studies. The dictionary contains a collection of non-descriptive words that have no bearing on the document. Stopword removal will consider the presence of words found in the Indonesian sentiment lexicon, negation words, and conjunctions, as these words will influence feature extraction. In contrast, Saputra et al. [13] did not delete words from Indonesian sentiment lexicons, negative words, and conjunction words, even if they are included in the stopword, because they affect the feature extraction process.

In addition to stopword removal, the library can assist with the following pre-processing: Normalization, Stemming, and Tokenization. Several studies [8], [10], [20] used the NLTK library for tokenization. Natural Language Toolkit (NLTK) is a natural language processing (NLP) library written in Python that can perform both symbolic and statistical processing. NLTK can display graphical demonstrations but requires sample data to process. This NLTK fully supports the fields of empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK can perform semantic reasoning, classification, tokenization, stemming, tagging, and parsing functions. The nltk.tokenize module was required for the tokenization step. It parses a string of characters and returns the syllables from a single word.

According to [8], [10], [18], [20], [21], [24], the Sastrawi Library can help with the stemming process. Python Sastrawi is a straightforward library that converts Indonesian affixed words to their basic form. This library is widely used in many Bahasa Indonesia text analytics studies. This procedure seeks to increase the likelihood that similar words with similar root words are counted as one. However, some studies did not use stemming because it can alter the meaning of the entire sentence [7].

The normalization process can be done manually or with the assistance of a pre-existing wordlist derived from previous research. As demonstrated in [19], abbreviations and misspellings in tweets were normalized by using a manually created Indonesian typography dictionary. Other studies [10], [23] used "kamus alay" or Colloquial Indonesian Lexicon [26] to perform normalization. This lexicon was created by collecting 3,592 unique colloquial words, also known as "bahasa alay," and manually annotating them with the normalized form. Another study [15] converted Indonesian slang words into standard words with the help of a GitHub dictionary by Louis Owen. Aziz's research data were used for normalization in [13], [17]. Several studies [16], [18] did text normalization using dictionaries obtained from a combination of dictionaries from previous works. This dictionary is a continuous, collaborative work based on Indonesian language research [26]–[28]. The text normalization dictionary contains 11,034 terms that have been normalized. In addition to lemmatization, the dictionary helps with Indonesian abbreviations, slang, misspelled words, and even the names of political figures. As a result, the normalized form frequently contains more than one word.

Aside from the previously mentioned pre-processing, data cleaning is a process that is almost universally used in research. Cleaning is a process that removes unnecessary elements from data, but it has a different meaning in each study. This is determined by the type of data used and the data to be extracted. Cleaning is performed on Twitter data to remove hashtags beginning with the # symbol, mentions beginning with the @ symbol, links, usernames, retweets beginning with RT, HTML tags, scripts, numbers, images, and whitespace. Data from Google Play is cleaned by deleting the application's name used as a keyword while scraping [7]. Some studies [8], [24] defined cleaning as the process of removing duplicates. Stopword removal is included in [19] cleaning process. Other research [10], [11], [14], [18], [22] included the removal of punctuation during the cleaning process.

Meanwhile, in some studies [13], [17], punctuation was not removed. This is because punctuation was used to separate clauses. The symbols not deleted in this study were based on a reference from a previous study's punctuation list. Another study [19] kept the original dataset in a separate file because orthography information (such as capital characters and punctuation) can be used to detect emotion.

4 Conclusion

Most sentiment analysis research utilizing Inset in the previous five years focused on the health domain, notably the COVID-19 pandemic. Discussions regarding the COVID-19 pandemic include the issue of COVID-19 in the community, health protocols, vaccine developments, and government strategies in dealing with COVID-19, such as lockdown, PSBB, new normal, and quarantine during a pandemic. This is in line with society's current state, which is still recovering from the pandemic's effects. Twitter is most likely the appropriate data source because the Inset dictionary was also created using words from Twitter. This explains why the majority of the data came from Twitter. Twitter data crawling can be accomplished in various ways, including using the Twitter API and the python library. Tweepy, Twint, and GetOldTweets are three Python libraries that can assist us in crawling Twitter data. The common pre-processing techniques are stopword removal, tokenization, case folding, and stemming. Several libraries can assist with the pre-processing method, such as the NLTK library for tokenization and the Sastrawi library for stemming. Furthermore, previous research data can be used as a reference in pre-processing. Other pre-processing techniques must be adjusted to the dataset in question. This study only discussed the data and pre-processing in lexicon-based sentiment analysis using an Inset dictionary. In further work, it is expected to discuss in more detail the lexicon method used and also its accuracy. This will undoubtedly provide a more thorough picture of the present trend of lexicon-based sentiment analysis using the Inset dictionary.

References

[1] Y. Fauziah, B. Yuwono, and A. S. Aribowo, “Lexicon based sentiment analysis in Indonesia languages : A systematic literature review,” *RSF Conf. Ser. Eng. Technol.*, vol. 1, no. 1, pp. 364–367, 2021.

[2] T. Eng, M. R. Ibn Nawab, and K. M. Shahiduzzaman, “Improving accuracy of the sentence-level lexicon-based sentiment analysis using machine learning,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 57–69, 2021, doi: 10.32628/cseit21717.

[3] F. Hemmatian and M. K. Sohrabi, “A survey on classification techniques for opinion mining and sentiment analysis,” *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1495–1545, 2019.

[4] G. N. Alemneh, A. Rauber, and S. Atnafu, “Dictionary based amharic sentiment lexicon generation,” in *International Conference on Information and Communication Technology for Development for Africa*, 2019, pp. 311–326.

[5] M. Darwich, S. A. Mohd, N. Omar, and N. A. Osman, “Corpus-based techniques for sentiment lexicon generation: A Review.,” *J. Digit. Inf. Manag.*, vol. 17, no. 5, p. 296, 2019.

[6] I. Guellil, A. Adeel, F. Azouaou, and A. Hussain, “Sentialg: Automated corpus annotation for algerian sentiment analysis,” in *International conference on brain inspired cognitive systems*, 2018, pp. 557–567.

[7] K. S. Nugroho, A. Y. Sukmadewa, F. A. Bachtiar, and N. Yudistira, “BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews,” *arXiv*, pp. 1–10, 2020.

[8] A. D. Widiantoro, A. Wibowo, and B. Harnadi, “User sentiment analysis in the fintech OVO

review based on the lexicon method," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 2021, pp. 1–4.

[9] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," *Proc. 2017 Int. Conf. Asian Lang. Process. IALP 2017*, vol. 2018-Janua, pp. 391–394, 2018, doi: 10.1109/IALP.2017.8300625.

[10] R. Jayapermana, A. Aradea, and N. I. Kurniati, "Implementation of stacking ensemble classifier for multi-class classification of COVID-19 vaccines topics on Twitter," *Sci. J. Informatics*, vol. 9, no. 1, pp. 8–15, 2022.

[11] D. D. Heriswan, Y. A. Sari, and M. T. Furqon, "Clustering Public Opinions Related to Quarantine during Covid-19 on Twitter Using K-DENCLUE Algorithms," in *6th International Conference on Sustainable Information Engineering and Technology 2021 (SIET '21)*, 2021, pp. 252–257.

[12] T. D. Dikiyanti, A. M. Rukmi, and M. I. Irawan, "Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm," *J. Phys. Conf. Ser.*, vol. 1821, no. 1, 2021, doi: 10.1088/1742-6596/1821/1/012054.

[13] F. T. Saputra, S. H. Wijaya, Y. Nurhadryani, and Defina, "Lexicon addition effect on Lexicon-Based of Indonesian sentiment analysis on Twitter," in *Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020*, 2020, pp. 136–141, doi: 10.1109/ICIMCIS51567.2020.9354269.

[14] R. L. Mustofa and B. Prasetiyo, "Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter," *ICMSE 2020. J. Phys. Conf. Ser.*, 2020, doi: 10.1088/1742-6596/1918/4/042155.

[15] A. Muhaddisi, B. N. Prastowo, and D. U. K. Putri, "Sentiment analysis with sarcasm detection on politician's Instagram," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 4, pp. 349–358, 2021, [Online]. Available: https://journal.ugm.ac.id/ijccs/article/view/66375.

[16] R. Suryadikara, S. Verberne, and ..., "False news classification and dissemination: the case of the 2019 Indonesian presidential election," 2020, [Online]. Available: https://scholarlypublications.universiteitleiden.nl/access/item%3A3070912/view.

[17] F. T. Saputra and Y. Nurhadryani, "Analysis of Indonesian sentiments using Indonesian sentiment lexicon by considering denial," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 361–366.

[18] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 46–57.

[19] M. S. Saputri, R. Mahendra, and ..., "Emotion classification on indonesian twitter dataset," in *International Conference on Asian Language Processing (IALP)*, 2018, pp. 90–95, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8629262/.

[20] S. Amalina, "Sentiresearch: lexicon-based web application for Indonesian sentiment analysis," dspace.uii.ac.id, 2020.

[21] R. Firdaus, I. Asror, and A. Herdiani, "Lexicon-based sentiment analysis of Indonesian language student feedback evaluation," *Indones. J. Comput.*, vol. 6, no. 1, pp. 1–12, 2021, [Online]. Available: http://socj.telkomuniversity.ac.id/ojs/index.php/indojc/article/view/408.

[22] R. H. Muhammadi, T. G. Laksana, and ..., "Combination of Support Vector Machine and lexicon-based algorithm in Twitter sentiment analysis," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 8, no. 1, 2022, [Online]. Available:

https://journals.ums.ac.id/index.php/khif/article/view/15213.

[23] M. C. Rahmadan, A. N. Hidayanto, and ..., “Sentiment analysis and topic modelling using the LDA method related to the flood disaster in Jakarta on Twitter,” in *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2020, pp. 126–130, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9354320/.

[24] F. H. Rachman and B. S. Rintyarna, “Sentiment analysis of Madura tourism in new normal era using Text Blob and KNN with Hyperparameter Tuning,” in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 2022, pp. 23–27.

[25] Y. Mahendradhata *et al.*, “The Republic of Indonesia health system review,” 2017.

[26] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, and A. Jamal, “Colloquial indonesian lexicon,” in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 226–229.

[27] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, “Hate speech detection in the Indonesian language: A dataset and preliminary study,” in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, pp. 233–238.

[28] M. O. Ibrohim and I. Budi, “A dataset and preliminaries study for abusive language detection in Indonesian social media,” *Procedia Comput. Sci.*, vol. 135, pp. 222–229, 2018.