Sentiment Analysis of Air Pollution on Social Media: Systematic Literature Review

Yandi Dwi Permana*, Abdul Gofur, Indra Budi, Aris Budi Santoso, Prabu Kresna Putra
Program Magister Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Indonesia,
Jl. Salemba Raya No.4, DKI Jakarta, Indonesia, 10430
*e-mail: yandi.dwi@office.ui.ac.id

(***received***: 28 November 2023, ***revised***: 20 March 2024, ***accepted***: 22 April 2024)

Abstract

The need for a healthy and pollution-free environment is the basis of the problem that this study examines. Social media has become an integral aspect of daily existence for the majority engaged in the digital realm. It enables individuals from various backgrounds to utilize these platforms to stay updated on the latest information, such as the current state of pollution in Jakarta. This research explores the attitudes of social media users regarding their perspectives on air pollution in Jakarta. The method used includes conducting a Systematic Literature Review of academic papers released from 2020 to 2023. The results of this research can unveil the types of social media platforms utilized, the quantity of datasets, the procedures for data collection, data preprocessing techniques, and the commonly employed methods in sentiment analysis studies concerning the subject of air pollution.

Keyword: air pollution, sentiment analysis, systematic literature review

1 Introduction

In the present age of globalization, individuals frequently move between different locations, emphasizing the urgent requirement for a pollution-free and comfortable environment. The circumstances of a region regarding disasters undergo frequent and dynamic changes, highlighting the need for a reporting system that is systematic, contextual, and real-time to obtain swift and precise regional evaluations. Social media presents itself as a potential solution for the envisaged reporting system [1].

Social media is one of the tools used by netizens to access, share, and discuss recent issues. Through sentiment analysis on social media, we can understand how people describe and express their perceptions of air pollution conditions in Jakarta, whether positively, negatively, or neutrally. Not all opinions can be used for analysis, so a sorting process, as in research, must be carried out. For example, in research [2], only positive and negative opinions are used to analyze the impact of opinions on social media. Currently, programming languages for data mining analysis have seen significant development. In research [3], R programming is used to analyze sentiment data from Twitter, while other studies employ various programming languages.

The objective of this research is to illustrate the commonly employed techniques for gathering, preprocessing, and categorizing social media data concerning the trends associated with air pollution in Jakarta. Additionally, the study seeks to present an overview of sentiment data derived from various studies that have analyzed social media sentiment on the topic. The methodology employed in this study involves utilizing the Systematic Literature Review approach to examine existing literature and address the identified issue.

2 Literature Review

Air pollution stands out as a critical environmental issue in the contemporary era. It emanates from diverse sources, encompassing motor vehicles, industrial emissions, and open burning. The repercussions of air pollution extend to adverse effects on air quality, human well-being, and the ecosystem. Deteriorating air quality can precipitate various health issues, ranging from respiratory and cardiovascular diseases to cancer. Furthermore, pre-existing health conditions like asthma and heart disease can exacerbate due to air pollution. The environmental ramifications include contributions to global warming, acid rain, and the depletion of the ozone layer [4].

Sentiment analysis, also known as opinion mining and trend analysis, aims to discern the emotional polarity by differentiating between negative and positive sentiments while also measuring the intensity of these emotions. Notably, dictionary-based sentiment analysis represents a method that doesn't necessitate supervised training, making it a prevalent and straightforward approach. In essence, sentiment analysis or opinion mining involves discerning attitudes directed toward objects or people [5]. Sentiment analysis holds three distinct components: positive sentiment, negative sentiment, and neutral sentiment. It can also be intricately dissected to identify the specific individual or group responsible for generating positive or negative sentiments. Dictionary-based sentiment analysis involves computing sentiment values by referencing a labeled sentiment dictionary.

The Systematic Literature Review method is employed to systematically recognize, assess, and interpret all pertinent and chosen research within the specific phenomenological subject of interest. This process addresses a set of questions related to the relevant research. Systematic Literature Review is a form of secondary research that uses a well-defined methodology. The Systematic Literature Review method can be reviewed and determined systematically, in each process following the predetermined steps. The main goal of a Systematic Literature Review is to provide a thorough compilation of all published research related to a specific field of study [6].

3 Research Method

The stages of the Systematic Literature Review conducted in this research were adopted from the Kitchenham method, versions 1.0 [7] and 2.3 [6]. There are three stages involved in Systematic Literature Review: planning, implementation, and reporting is illustrated in Figure 1.

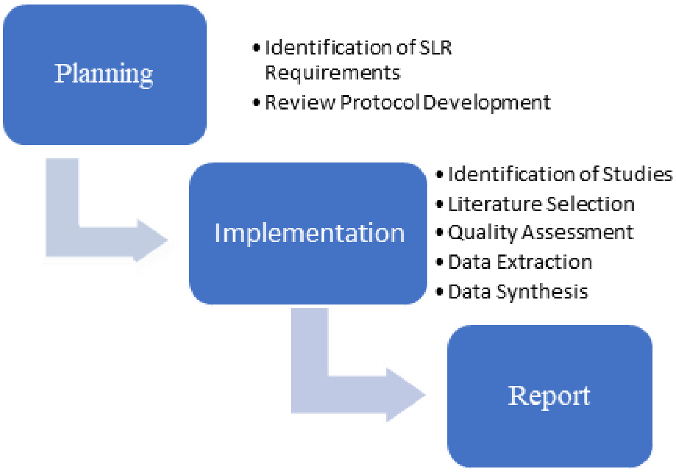

Figure 1. Systematic Literature Review (SLR) Methodology [7]

3.1 Planning

In the planning stage, we will discuss the identification of Systematic Literature Review requirements, which involve the need to understand public sentiment regarding air pollution in Jakarta, as well as the development of a review protocol that generates five research questions, namely:

RQ 1: Which social media platforms are used for sentiment analysis?
RQ 2: How many datasets are utilized for sentiment analysis?
RQ 3: What techniques are utilized for data collection in sentiment analysis?
RQ 4: Which techniques are utilized for preprocessing data in sentiment analysis?
RQ 5: What methods are employed in conducting sentiment analysis?

3.2 Implementation

In the implementation stage, we will conduct the identification of studies, literature selection, quality assessment, data extraction, and data synthesis.

3.2.1 Identification of Studies

In this study, literature search was conducted using the keyword "sentiment analysis" AND (pollution OR environmental OR climate OR plastic OR waste)," sourced from the Scopus, ScienceDirect, and IEEE Xplore databases, which included journals and conference papers/proceedings. Afterward, additional refinement was undertaken, directed by specific criteria for inclusion and exclusion. The criteria for inclusion in this study are outlined as follows:

1. Journals and conference papers collected are research published from 2020 to 2023
2. The journals and conference papers are authored in the English language
3. The study revolves around sentiment analysis concerning environmental concerns, with a specific focus on air pollution

Meanwhile, the criteria for excluding content in this study are outlined as follows:

1. Journals and conference papers published outside the range of 2020 to 2023
2. Journals and conference papers not written in English
3. Research unrelated to sentiment analysis of the environment

3.2.2 Literature Selection

Based on the search, a total of 15 journals or conference papers were found, and the search process was divided into three stages. The inittiation stage involved searching according to the keywords, inclusion criteria, and exclusion criteria as previously described, resulting in 208 literature. The first stage was carried out by examining the titles and abstracts, which yielded 52 literature. The second stage involved reading the entire journals, resulting in 15 literature. The literature selection results is illustrated in Table 1.

Table 1. Literature Selection

Source	Initiation Stage (Based on the search results)	Stage 1 (Title and Abstract Selection)	Stage 2 (Full Text Selection)
Scopus	169	39	11
ScienceDirect	27	7	1
IEEE Xplore	12	6	3
Total	**208**	**52**	**15**

3.2.3 Quality Assessment

Quality assessment is carried out to uphold the standard of the chosen literature. Due to the limited number of literature related to sentiment analysis of air pollution or environmental themes, it can be quite challenging to assess the quality of these literature pieces. The criteria for quality assessment is illustrated in Table 2.

Table 2. Quality Assessment Checklist

Checklist	Questions
C1	Are the research objectives explicitly outlined in the literature?
C2	Does the literature include a comprehensive review, background, and contextualization of the research?
C3	Is there an explanation of the data collection methods employed in the literature?
C4	Does the literature explain the data preprocessing methods used?

C5	Does the literature describe the methods used?
C6	Does the literature provide conclusions that are pertinent to the research objectives or questions?
C7	Does the literature propose future endeavors or suggest enhancements for upcoming work?
C8	Is the literature indexed in Scopus (Q1/Q2/Q3/Q4/unindexed)?

3.2.4 Data Extraction

During this phase, information is gathered from the examined literature, encompassing details such as the paper's publication year, the social media platform utilized, the dataset employed, the techniques for data collection, the applied data preprocessing methods, and the approaches utilized for sentiment analysis. This information is then compiled and structured in a spreadsheet document.

3.2.5 Data Synthesis

During this phase, the literature is chosen based on its potential to contribute to addressing the research queries and compliance with the specified inclusion and exclusion criteria. Following the review, 15 literatures were identified for further examination.

4 Results and Analysis

This research has collected a total of 15 relevant previous literature related to the title. These selected literature will undergo additional scrutiny pertaining to the five topics delineated in the previously mentioned research questions: social media, quantity of datasets, data collection methods, data preprocessing, and sentiment analysis approaches.

4.1 Media Sosial

Out of the 15 analyzed literature, 11 of them [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] utilized the social media platform Twitter for sentiment analysis, 1 literature [19] used Weibo, which is a commonly used social media platform in China, and 2 literature [20], [21] used government websites. Furthermore, there is one literature [22] that utilized three social media platforms, namely Twitter, Reddit, and Youtube. The utilization of social media platforms in sentiment analysis is illustrated in Figure 2.

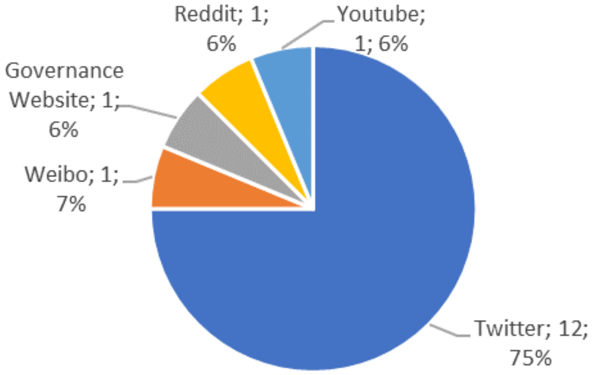

Figure 2. Media Social Utilization

4.2 Dataset

The average number of datasets used in the research is 94,110 datasets, with the smallest dataset comprising 989 datasets in the study [15] and the largest dataset consisting of 795,065 datasets in the study [18]. Additionally, there are three studies that do not mention the number of datasets processed [8], [12], [20].

4.3 Data Collection

In general, sentiment analysis studies concerning data collection methods can be categorized into two groups. The initial category encompasses sentiment analysis research that relies on existing public datasets, while the second category involves studies that conduct their own data collection. Data can consist of text, images, audio or video.

The data collection found in the 15 reviewed literatures can be categorized into three groups, namely data collection techniques, data source/format, and tools. In accordance with the mentioned categories, data collection used are: 1) Technique: Crawling Data by authors [20], [21]; 2) Data Source/Format: Twitter API by authors [8], [9], [11], [12], [14], [15], [18], Weibo API by authors [19], Tweepy by authors [16], [17], and Reddit API by authors [22]; 3) Tools: Python Library by authors [10], [16], SNScrape by authors [10], and R vosonSML package by authors [22]. Figure 3 displays the data collection alongside the corresponding research counts.

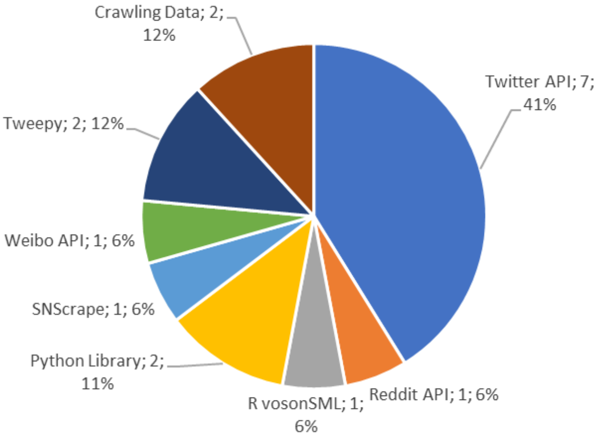

Figure 3. Data Collection

4.4 Data Preprocessing

Data preprocessing plays a crucial role in sentiment analysis, particularly when dealing with unstructured data. Classification algorithms generally do not accept unstructured data directly as input, as they are designed to process numerical data. Additionally, algorithms may struggle to comprehend the diverse writing styles, languages, and even individual writing approaches.

The first step in data preprocessing involves cleaning the data. During this cleaning stage, extraneous characters, words, phrases, and even entire tweets are eliminated from the dataset. The data preprocessing techniques utilized in sentiment analysis, as identified in the 15 literature pieces, are detailed in Table 3.

Table 3. Data Prepocessing Technique

Literature	Data Preprocessing Technique
[8], [10], [11], [14]	Tokenization
[8], [10], [11], [15], [20], [22]	Stop-word removal
[8], [10], [13]–[15], [22]	Stemming
[8]	Feature extraction
[9]–[13], [17]–[20], [22]	Cleaning the punctuation, url, and emojis
[10], [15]	Case Folding
[10]	Bag of words
[10]	Labeling data

4.5 Methods

From the examination of 15 literature, various approaches were identified and can be categorized into five groups employed for sentiment analysis namely Classic, Dictionary-Based, Ensemble, Neural Network/Deep Learning, and Statistics. In accordance with the mentioned categories, the methods used are: 1) Classic: Logistic Regression (LR), Naïve Bayes (NB), and Support vector machine (SVM); 2) Dictionary Based: Valence Aware Dictionary and sEntiment Reasoner (VADER); 3) Ensemble: Gradient Boosting (GB) and Random Forest (RF); 3) Neural Network/Deep Learning: Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Long Short-Term Memory (Bi-LSTM), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and ResNet; 4) Statistic: Latent Dirichlet Allocation (LDA).

Furthermore, when considering the years in which the research was conducted, the category of methods most frequently employed over the last two years is Neural Network/Deep Learning, found in 7 literature [8], [9], [10], [13], [18], [20], [22] followed by the Classic category present in 6 literature [8], [10], [11], [12], [13], [22]. The methods used in sentiment analysis found in the 15 literature is illustrated in Table 4.

Table 4. Methods of Sentiment Analysis

Literature	Methods	Method Category	Year
[8], [9], [18]	BERT	Neural Network/Deep Learning	2023: [8], [9], [18]
[9], [17]	Bi-LSTM	Neural Network/Deep Learning	2021: [17] 2023: [9]
[10], [22]	CNN	Neural Network/Deep Learning	2023: [10], [22]
[21]	Dictionary Based	Dictionary Based	2021: [21]
[10], [22]	GB	Ensemble	2023: [10], [22]
[10]	LDA	Statistics	2023: [10]
[10], [22], [11]	LR	Classic	2022: [11] 2023: [10], [22]
[18], [20]	LSTM	Neural Network/Deep Learning	2023: [18], [20]
[8], [11], [12]	NB	Classic	2022: [11], [12] 2023: [8]
[13]	RF	Ensemble	2022: [13]
[13], [22]	RNN	Neural Network/Deep Learning	2022: [13] 2023: [22]
[18]	ResNet	Neural Network/Deep Learning	2023: [18]
[10], [11], [13], [15], [19], [22]	SVM	Classic	2021: [15], [19] 2022: [11], [13] 2023: [10], [22]
[16], [18]	VADER	Dictionary Based	2023: [16], [18]

5 Conclusion

This research compiles various sentiment analyses related to environmental topics, particularly air pollution. Literature search was conducted by reviewing the titles, abstracts, and content of studies found in IEEE Xplore, ScienceDirect, and Scopus, resulting in 15 literature to be further analyzed. The examination is carried out with regard to the choice of social media, the volume of datasets utilized, data collection methodologies, data preprocessing, and the approaches employed in sentiment analysis.

The choice of social media platform most commonly used is Twitter, which appears in 12 literature. The number of datasets used, as analyzed in the 15 literature, ranges from 989 datasets to 795,065 datasets. For data collection, the tool frequently used is the Twitter API, used in 7 literature. After data collection, data cleansing and preprocessing are necessary steps. The data preprocessing method used frequently used is the Cleaning the punctuation, url, and emojis, used in 10 literature. Meanwhile, within the pool of 15 literature items, the Support Vector Machine emerged as the most frequently utilized method, implemented by six of the studies.

This research has not yet provided a detailed discussion of the methods used. Hence, additional investigation is necessary to offer a thorough overview of the methodologies and the variables or parameters utilized. This will be beneficial for readers who wish to delve deeper into sentiment analysis and gain a better understanding of the techniques involved.

Reference

[1] P. Lei, G. Marfia, G. Pau, and R. Tse, "Can We Monitor the Natural Environment Analyzing Online Social Network Posts? A literature review," *Online Soc Netw Media*, vol. 5, pp. 51–60, 2018.

[2] Y. Setiowati and F. Setyorini, "Service Extraction and Sentiment Analysis to Indicate Hotel Service Quality in Yogyakarta based on User Opinion," in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 2018, pp. 427–432.

[3] S. Saini, R. Punhani, R. Bathla, and V. K. Shukla, "Sentiment Analysis on Twitter Data using R," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, IEEE, 2019, pp. 68–72.

[4] Kementerian Kesehatan Republik Indonesia, "Penting Pahami Ancaman Polusi Udara pada Kesehatan." Accessed: Sep. 23, 2023. [Online]. Available: https://ayosehat.kemkes.go.id/penting-pahami-ancaman-polusi-udara-pada-kesehatan

[5] P. Beineke, T. Hastie, and S. Vaithyanathan, "The Sentimental factor: Improving Review Classification via Human-Provided Information," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004, pp. 263–270.

[6] B. Kitchenham and S. Charters, "Guidelines for Performing Systematic Literature Reviews in Software Engineering Version 2.3," *Engineering*, vol. 45, no. 4ve, p. 1051, 2007.

[7] B. Kitchenham, "Procedures for Performing Systematic Reviews," *Keele, UK, Keele University*, vol. 33, no. TR/SE-0401, p. 28, 2004, doi: 10.1.1.122.3308.

[8] M. Lydiri, Y. El Mourabit, Y. El Habouz, and M. Fakir, "A performant Deep Learning Model for Sentiment Analysis of Climate Change," *Soc Netw Anal Min*, vol. 13, no. 1, Dec. 2023, doi: 10.1007/s13278-022-01014-3.

[9] A. Upadhyaya, M. Fisichella, and W. Nejdl, "Towards Sentiment and Temporal Aided Stance Detection of Climate Change Tweets," *Inf Process Manag*, vol. 60, no. 4, Jul. 2023, doi: 10.1016/j.ipm.2023.103325.

[10] N. Ashari, M. Z. Mifta Al Firdaus, I. Budi, A. B. Santoso, and P. Kresna Putra, "Analyzing Public Opinion on Electrical Vehicles in Indonesia using Sentiment Analysis and Topic Modeling," in *ICCoSITE 2023 - International Conference on Computer Science, Information Technology and Engineering: Digital Transformation Strategy in Facing the VUCA and TUNA Era*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 461–465. doi: 10.1109/ICCoSITE57641.2023.10127834.

[11] N. M. Sham and A. Mohamed, "Climate Change Sentiment Analysis using Lexicon, Machine Learning and Hybrid Approaches," *Sustainability (Switzerland)*, vol. 14, no. 8, Apr. 2022, doi: 10.3390/su14084723.

[12] A. Ridwan, H. H. Nuha, and R. Dharayani, "Sentiment Analysis of Floods on Twitter Social Media using the Naive Bayes Classifier Method with the N-Gram Feature," in *2022 International Conference on Data Science and Its Applications, ICoDSA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 114–118. doi: 10.1109/ICoDSA55874.2022.9862827.

[13] M. I. Alhari, O. N. Pratiwi, and M. Lubis, "Sentiment Analysis of The Public Perspective Electric Cars in Indonesia using Support Vector Machine Algorithm," in *2022 International Conference of*

Science and Information Technology in Smart Administration, ICSINTESA 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 155–160. doi: 10.1109/ICSINTESA56431.2022.10041604.

[14] S. Srivastava, J. P. Singh, and D. Manga, *Time and Domain Specific Twitter Data Mining for Plastic Ban based on Public Opinion*. 2020.

[15] Z. Indra, A. Setiawan, and Y. Jusman, "Implementation of Machine Learning for Sentiment Analysis of Social and Political Orientation in Pekanbaru City," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Feb. 2021. doi: 10.1088/1742-6596/1803/1/012032.

[16] A. Madjar, I. Gjorshoska, J. Prodanova, A. Dedinec, and L. Kocarev, "Western Balkan Societies' Awareness of Air Pollution. Estimations using Natural Language Processing Techniques," *Ecol Inform*, vol. 75, Jul. 2023, doi: 10.1016/j.ecoinf.2023.102097.

[17] T. Thukral, A. Varshney, and V. Gaur, "Intensity Quantification of Public Opinion and Emotion Analysis on Climate Change," *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, no. 83, pp. 1351–1366, Oct. 2021, doi: 10.19101/IJATEE.2021.874417.

[18] L. Bryan-Smith, J. Godsall, F. George, K. Egode, N. Dethlefs, and D. Parsons, "Real-Time Social Media Sentiment Analysis for Rapid Impact Assessment of Floods," *Comput Geosci*, vol. 178, Sep. 2023, doi: 10.1016/j.cageo.2023.105405.

[19] S. Shan, X. Ju, Y. Wei, and Z. Wang, "Effects of PM2.5 on People's Emotion: A Case Study of Weibo (Chinese Twitter) in Beijing," *Int J Environ Res Public Health*, vol. 18, no. 10, May 2021, doi: 10.3390/ijerph18105422.

[20] S. Puppala, I. Hossain, and S. Talukder, "Machine Learning and Sentiment Analysis for Predicting Environmental Lead Toxicity in Children at the ZIP Code Level," in *2023 IEEE 2nd International Conference on AI in Cybersecurity, ICAIC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICAIC57335.2023.10044177.

[21] Y. Sun, F. Jin, Y. Zheng, M. Ji, and H. Wang, "A New Indicator to Assess Public Perception of Air Pollution based on Complaint Data," *Applied Sciences (Switzerland)*, vol. 11, no. 4, pp. 1–18, Feb. 2021, doi: 10.3390/app11041894.

[22] K. McGarry, "Analyzing Social Media Data using Sentiment Mining and Bigram Analysis for the Recommendation of YouTube Videos," *Information (Switzerland)*, vol. 14, no. 7, Jul. 2023, doi: 10.3390/info14070408.