

Sentiment Analysis of youtube Comments on "Cerdas Cermat Empat Pilar MPR RI "Using Lexicon-Based Sentiment Analysis"

Nabil Astian^{1*}, Sofyh Zack Savira², Evi Morina Br.Tarigan³, Natalia Situmeang⁴

^{1 2 3 4} Department of Informatics, Universitas Satya Terra Bhinneka, Medan, Indonesia

ABSTRACT

The rapid growth of social media has transformed digital communication, enabling citizens to express opinions openly on various public issues. This study examines YouTube user comments on the video "Cerdas Cermat Empat Pilar MPR RI West Kalimantan Province" to understand public sentiment toward civic education content. Manual analysis of large-scale YouTube comment data is inefficient and prone to subjectivity. This study addresses the question: how can Lexicon-Based Sentiment Analysis effectively classify public opinion in Indonesian-language YouTube comments? This study applies a lexicon-based rule-based approach without training data requirements, providing an accessible alternative for sentiment analysis on medium-scale Indonesian-language datasets with full preprocessing pipeline integration. A total of 1,709 comments were collected via web scraping using the YouTube Comment Downloader library on Google Colaboratory. Data preprocessing included case folding, text cleaning, stopword removal, Sastrawi stemming, and negation handling. Sentiment scoring was performed using a customized Indonesian sentiment lexicon. Neutral sentiment dominated with 857 comments (50.15%), followed by negative sentiment at 516 comments (30.19%), and positive sentiment at 336 comments (19.66%). Classification errors were identified in negation and sarcasm processing. Lexicon-Based Sentiment Analysis is a practical and efficient method for public opinion analysis on YouTube. The dominance of neutral comments indicates informative discussion patterns, while high negative proportions reflect a tendency toward public criticism. Future work should integrate machine learning comparisons and expanded lexicon resources.

ARTICLE INFO

Keywords:
Sentiment Analysis, YouTube, Text Mining, Lexicon-Based, Opinion Mining

Histori:
Received: 02 July 2026
Accepted: 03 July 2026
Published: 29 July 2026

* Corresponding Author at Department of Informatics, Universitas Satya Terra Bhinneka, Medan, Indonesia.
E-mail address: 2303310344@students.satyaterrabhinneka.ac.id

1. INTRODUCTION

The rapid advancement of digital technology has significantly transformed the way people communicate and exchange information through social media platforms (Lindgren 2021). Social media has evolved beyond a medium for sharing information into an interactive digital environment where users actively express opinions, experiences, and perceptions regarding various social, political, educational, and cultural issues. Among the available platforms, YouTube has become one of the world's largest video-sharing platforms, enabling users not only to watch content but also to participate in discussions through the comment section (Alhujaili and Yafooz 2021). Consequently, YouTube comments constitute a valuable source of user-generated textual data that reflects the sentiments expressed by users toward specific content.

The large volume of comments generated on YouTube makes manual sentiment analysis inefficient, time-consuming, and susceptible to subjective interpretation. To overcome this limitation, sentiment analysis has become one of the most widely adopted applications of Natural Language Processing (NLP) for automatically identifying sentiment polarity from textual data (Fudholi, Rahaningsih, and Dana 2024). By classifying comments into positive, negative, and neutral categories, sentiment analysis enables researchers to systematically analyze user responses and identify general sentiment patterns within large collections of comments.

Recent sentiment analysis studies have demonstrated significant progress across various domains. Most previous studies have focused on Twitter data, e-commerce product reviews, online marketplace reviews, hotel reviews, or mobile application reviews using supervised Machine Learning and Deep Learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and IndoBERT (Semary et al. 2024) (Moiceanu 2025). Although these approaches generally achieve high classification performance, they require manually labeled datasets, extensive training processes, and considerable computational resources (Rahman et al. 2021). In contrast, the Lexicon-Based Sentiment Analysis approach offers a practical alternative because it does not require labeled training data, is computationally efficient, and can be effectively implemented on small- and medium-sized Indonesian-language datasets (Hill and Fresneda 2025).

Several previous studies have also analyzed sentiment on YouTube. However, most of these studies focused on entertainment content, political issues, product reviews, or application reviews, while relatively limited attention has been given to Indonesian-language comments discussing civic education content (Sulaiman et al. 2025). Furthermore, many lexicon-based studies primarily emphasize sentiment classification results without providing a comprehensive explanation of the preprocessing workflow. Important preprocessing stages, including informal word normalization, emoji handling, stopword removal, stemming, negation handling, and lexicon customization, are frequently described only briefly, even though these processes

substantially influence sentiment classification performance for Indonesian social media text. (Sulaiman et al. 2025)

Based on these observations, three research gaps can be identified. First, studies investigating sentiment expressed in Indonesian-language YouTube comments related to civic education remain relatively limited compared with studies conducted on Twitter, e-commerce platforms, and entertainment content. Second, previous studies have generally paid limited attention to documenting an integrated preprocessing pipeline capable of handling the linguistic characteristics of Indonesian YouTube comments, including informal expressions, abbreviations, negation, and text normalization. Third, although Lexicon-Based Sentiment Analysis has been widely implemented, only a limited number of studies have reported the use of a customized Indonesian sentiment lexicon specifically adapted to the characteristics of YouTube comments while providing a transparent preprocessing workflow.

Therefore, this study proposes the application of Lexicon-Based Sentiment Analysis to analyze Indonesian-language comments collected from the YouTube video *"Cerdas Cermat Empat Pilar MPR RI Provinsi Kalimantan Barat."* (CNN Indonesia 2026) The proposed framework integrates several preprocessing stages, including case folding, text cleaning, informal word normalization, stopword removal, Sastrawi stemming, emoji removal, and negation handling before performing sentiment scoring using a customized Indonesian sentiment lexicon. This integrated preprocessing pipeline is intended to improve text consistency and support more reliable sentiment classification.

The novelty of this research lies not merely in applying the Lexicon-Based Sentiment Analysis method, but in presenting a comprehensive framework for analyzing Indonesian-language YouTube comments on civic education content through the integration of a customized Indonesian sentiment lexicon with a complete preprocessing workflow specifically designed to address the linguistic characteristics of social media text. In addition, this study provides a detailed analysis of sentiment distribution together with a discussion of linguistic challenges encountered during lexicon-based classification, including informal language, negation, contextual ambiguity, and sarcasm, thereby offering empirical evidence for the implementation of lexicon-based sentiment analysis on Indonesian YouTube comment datasets (Kumar, Khan, and Chang 2025).

Accordingly, the objective of this study is to classify sentiment expressed in the collected Indonesian-language YouTube comments using the Lexicon-Based Sentiment Analysis approach and to analyze the distribution of positive, negative, and neutral sentiments generated from the collected dataset. The findings are expected to contribute to the development of Indonesian-language sentiment analysis research by providing a transparent and reproducible lexicon-based analysis framework while enriching empirical studies on sentiment expressed in YouTube comments related to civic education content.

2. LITERATURE REVIEW

2.1. Social Media and YouTube as a Source of Public Opinion

Social media has become one of the most influential digital communication platforms in the modern era, enabling users to create, share, and exchange information in various formats, including text, images, audio, and video (Srg and Usiono 2024). Beyond facilitating interpersonal communication, social media has evolved into a virtual public space where individuals actively express opinions, experiences, and perceptions regarding social, political, economic, and cultural issues (Setiadarma et al. 2024). The rapid growth of user-generated content has resulted in an enormous volume of textual data that reflects public attitudes and behavioral trends (Rahmatullah 2021).

Among the various social media platforms, YouTube is recognized as the world's largest video-sharing platform, owned by Google LLC (Fauzi, Effendi, and Dity 2024). Besides allowing users to upload and watch videos, YouTube provides an interactive comment section where viewers can express opinions, provide feedback, and engage in discussions related to video content. These comments constitute valuable textual data that can be analyzed to understand public sentiment toward specific topics, making YouTube an important data source for sentiment analysis research (Widagdo and Saputro 2024).

2.2. Text Mining and Sentiment Analysis

Text mining is a computational approach designed to extract meaningful information and hidden knowledge from unstructured textual data (Hassani et al. n.d.). It integrates techniques from natural language processing (NLP), computational linguistics, statistics, and machine learning to transform raw text into structured information suitable for systematic analysis. Through preprocessing and analytical procedures, text mining enables researchers to identify patterns, discover relationships among terms, and generate insights from large collections of textual documents.

One of the most prominent applications of text mining is sentiment analysis, also referred to as opinion mining. Sentiment analysis aims to identify, extract, and classify subjective opinions or emotions expressed in textual data into predefined sentiment categories, typically positive, negative, and neutral (Fudholi et al. 2024). This technique has been extensively applied in numerous domains, including customer satisfaction assessment, product review evaluation, social media monitoring, market research, and public opinion analysis. By automatically identifying sentiment polarity, sentiment analysis provides valuable insights into how individuals perceive a particular issue or event.

2.3. Lexicon-Based Sentiment Analysis and Sastrawi Stemming Algorithm

Among various sentiment classification techniques, the lexicon-based approach is one of the most widely adopted rule-based methods (Han et al. 2018). Instead of relying on supervised learning models, this approach utilizes a predefined sentiment lexicon in which each word has an associated positive or negative sentiment score. The overall sentiment of a document is determined by aggregating the scores of all sentiment-bearing words contained in the text and comparing the final score with predefined polarity thresholds. Because it does not require labeled training data, the lexicon-based approach is relatively simple to implement, computationally efficient, and particularly suitable for small- and medium-sized datasets.

Since this research analyzes comments written in Indonesian, an appropriate preprocessing technique is required to address the language's rich morphological structure. Therefore, the Sastrawi stemming algorithm is employed to convert inflected words into their corresponding root forms by removing prefixes, suffixes, infixes, and circumfixes. This normalization process reduces lexical variation and improves the consistency of textual representation, thereby increasing the accuracy of matching words with entries in the sentiment lexicon. Consequently, the combination of the lexicon-based approach and Sastrawi stemming enhances the effectiveness of sentiment classification by ensuring that semantically equivalent words are treated as identical during the analysis process (Fauziah, Yuwono, and Aribowo 2021).

3. METHOD, DATA, AND ANALYSIS

3.1. Research Design and Flow

This study employed a quantitative descriptive research design to analyze sentiment expressed in Indonesian-language YouTube comments using the Lexicon-Based Sentiment Analysis approach. The research was conducted through five sequential stages, namely research design, data collection, data preprocessing, lexicon-based sentiment analysis, and performance evaluation. The overall workflow was designed to transform raw YouTube comments into structured sentiment information through a transparent and reproducible process. Figure 1 illustrates the complete research framework adopted in this study.

The research data were obtained through a web scraping technique using the YouTube Comment Downloader library in the Google Colaboratory environment. Data collection was conducted on all publicly available comments from the selected YouTube video until the predetermined collection deadline. A total of 1,709 comments were successfully collected and used as the dataset for this study.

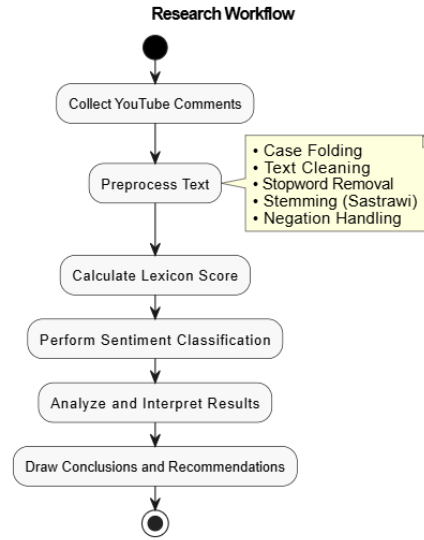


Figure 1. Research Flow

Before performing sentiment analysis, the collected comments underwent several preprocessing stages to improve data quality and ensure consistent text representation (Mola, Roma, and Widiastuti 2025). The first stage was case folding, in which all characters were converted into lowercase letters using the transformation:

$$T' = \text{Lowercase}(T) \quad (1)$$

The next stage was text cleaning, which removed irrelevant elements such as URLs, emojis, numeric characters, and special symbols from the comments. This process is represented by:

$$T'' = T' - \{URL, \text{Emoji}, \text{Numbers}, \text{Symbols}\} \quad (2)$$

Subsequently, stopwords removal was applied to eliminate common Indonesian words that carry little semantic information, such as conjunctions, prepositions, and articles. This operation is formulated as:

$$T''' = T'' - \text{Stopword} \quad (3)$$

After stopwords removal, the text was processed using the Sastrawi stemming algorithm to convert inflected words into their root forms, enabling more accurate matching with entries in the sentiment lexicon. The stemming process is expressed as:

$$\text{Stem}(T''') \rightarrow \text{Base Form} \quad (4)$$

To improve classification accuracy, negation handling was also implemented. This stage detects negation words such as *tidak*, *bukan*, and *jangan*, which may reverse the sentiment polarity of subsequent words. Proper handling of negation helps reduce sentiment misclassification caused by polarity inversion.

Finally, sentiment analysis was performed using a lexicon-based approach by calculating the sentiment score of each comment. The sentiment score was obtained by summing the weights of all sentiment-bearing words found in the lexicon. The calculation is defined as:

$$\text{Score}(d) = \sum_{i=1}^n w_i \quad (5)$$

where (Score(d)) represents the sentiment score of document (d), (w_i) denotes the sentiment weight of the (i)-th word in the lexicon, and (n) is the total number of sentiment words identified in the document. The final sentiment polarity was determined based on the resulting score, where positive values indicate positive sentiment, negative values indicate negative sentiment, and a score of zero indicates neutral sentiment.

3.2. Sentiment Classification

The dataset used in this study was collected from the YouTube video entitled “*Cerdas Cermat Empat Pilar MPR RI Provinsi Kalimantan Barat.*” Data collection was performed using the YouTube Comment Downloader library in the Google Colaboratory environment with Python. Publicly available comments were retrieved automatically through a web scraping process. The extracted data consisted of the username, comment text, publication time, and the number of likes associated with each comment.

To ensure the quality of the dataset, the collected comments were first examined before analysis. Duplicate comments, empty comments, promotional or spam comments, and comments written in languages other than Indonesian were excluded from the dataset. After the filtering process, a total of 1,709 valid Indonesian-language comments were retained for further analysis. The video URL [click here](#) (access date : June 25, 2026) ,Google Colaboratory environment, library version, and scraping configuration should also be reported to facilitate the reproducibility of this research.

Sentiment classification is performed based on the score values obtained from the calculation process, using the following threshold criteria:

Table 1. Sentiment Classification Criteria Based on Score Threshold

Score Condition	Sentiment Category
Score ≥ 2	Positive Sentiment
$-2 < \text{Score} < 2$	Neutral Sentiment
Score ≤ -2	Negative Sentiment

4. RESULT AND DISCUSSION

4.1. Descriptive Statistics of the Dataset

Based on the scraping process and sentiment classification carried out, the following data distribution was obtained as presented in Table 2:

Table 2. Distribution of Sentiment Classification Results

Sentiment Category	Number of Comments	Percentage (%)
Positive	336	19.66%
Negative	516	30.19%
Neutral	857	50.15%
Total	1,709	100.00%

4.2. Percentage Calculation of Sentiment Distribution

The percentage calculation for each sentiment category was performed using the following formula:

$$P = (n/N) \times 100\% \quad (6)$$

Based on this formula, the following calculations were obtained:

Positive Sentiment: $P = (336 / 1,709) \times 100\% = 19.66\%$

Negative Sentiment: $P = (516 / 1,709) \times 100\% = 30.19\%$

Neutral Sentiment: $P = (857 / 1,709) \times 100\% = 50.15\%$

4.3. Analysis and discussion

The sentiment analysis results show that neutral sentiment was the dominant category with 857 comments (50.15%), followed by negative sentiment with 516 comments (30.19%), and positive sentiment with 336 comments (19.66%). The high proportion of neutral comments indicates that most users used the comment section to exchange information, ask questions, or provide factual statements rather than express strong emotions.

Negative comments mainly criticized the judging panel's performance and frequently contained words such as *budeg*, *salah*, *kecewa*, and *parah*. This finding supports the negativity bias theory, which suggests that people are generally more likely to express negative experiences than positive ones on social media. Meanwhile, positive comments included words such as *hebat*, *bagus*, *keren*, and *mantap*, reflecting appreciation and support toward the video content.

Overall, the Lexicon-Based Sentiment Analysis method successfully classified Indonesian YouTube comments into positive, negative, and neutral categories. Although it is less capable of understanding context and sarcasm than machine learning methods, it remains practical because it does not require labeled training data, is computationally efficient, and is easy to implement for small- to medium-sized datasets.

5. CONCLUSION AND SUGGESTION

This study successfully implemented the Lexicon-Based Sentiment Analysis method to identify and classify YouTube user opinions toward the video "Cerdas Cermat Empat Pilar MPR RI West Kalimantan Province". A total of 1,709 comments were processed through a series of scraping, preprocessing, lexicon weighting, and sentiment classification stages.

Based on the conducted analysis, the following conclusions can be drawn:

- (1) The YouTube comment data collection process was successfully carried out using scraping techniques with the YouTube Comment Downloader library, yielding a total of 1,709 comments as the research corpus.
- (2) Preprocessing stages including case folding, text cleaning, stopword removal, Sastrawi stemming, and negation handling successfully prepared data for sentiment analysis.
- (3) Sentiment classification results showed: Neutral Sentiment = 857 comments (50.15%), Negative Sentiment = 516 comments (30.19%), and Positive Sentiment = 336 comments (19.66%).
- (4) The dominance of neutral sentiment indicates that YouTube comment sections are widely utilized as a medium for discussion and information exchange, while the high proportion of negative sentiment compared to positive reflects user tendencies to voice criticism.

(5) Lexicon-Based Sentiment Analysis proves applicable as an effective and efficient method for public opinion analysis on YouTube social media, despite limitations in processing complex linguistic phenomena such as negation, sarcasm, and irony.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used AI-assisted technologies, including Gemini and Google Translate, to support language editing, translation, and improvement of clarity and readability. The authors take full responsibility for the content of the manuscript and confirm that the use of these tools did not affect the originality, validity, or integrity of the research.

REFERENCE

Journal

- Alhujaili, R. F., and W. M. S. Yafooz. 2021. "Sentiment Analysis for Youtube Videos with User Comments: Review." Pp. 814–20 in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*.
- Fauzi, Unang, Erus Effendi, and Annisa Dity. 2024. "Optimasi Penggunaan Media Sosial Oleh Mahasiswa Hukum Ekonomi Syariah: Analisis Strategi Efektif Pada Platform YouTube." *Jurnal Komunikan* 3(1):55–73.
- Fauziah, Yuli, Bambang Yuwono, and Agus Sasmito Aribowo. 2021. "Lexicon Based Sentiment Analysis in Indonesia Languages : A Systematic Literature Review." 1(1):364–67. doi:10.31098/cset.v1i1.397.
- Fudholi, Luthfi Abdillah, Nining Rahaningsih, and Raditya Danar Dana. 2024. "Sentimen Analisis Perilaku Penggemar Coldplay Di Media Sosial Twitter Menggunakan Metode Naive Bayes." *JATI (Jurnal Mahasiswa Teknik Informatika)* 8(3):4150–59.
- Han, Hongyu, Yongshi Zhang, Jianpei Zhang, Jing Yang, and Xiaomei Zou. 2018. "Improving the Performance of Lexicon-Based Review Sentiment Analysis Method by Reducing Additional Introduced Sentiment Bias." 1–11.
- Hassani, Hossein, Christina Beneki, Stephan Unger, and Maedeh Taj Mazinani. n.d. "Text Mining in Big Data Analytics." 1–34. doi:10.3390/bdcc4010001.
- Hill, Chelsey H., and Jorge E. Fresneda. 2025. "The Wisdom of the Lexicon Crowds : Leveraging on Decades of Lexicon-Based Sentiment Analysis for Improved Results."
- Kumar, Mahander, Lal Khan, and Hsien-tsung Chang. 2025. "Evolving Techniques in Sentiment Analysis : A Comprehensive Review." 1–61. doi:10.7717/peerj-cs.2592.
- Lindgren, Simon. 2021. "Digital Media and Society."
- Moiceanu, Georgiana. 2025. "Developing a Deep Learning-Based Sentiment Analysis System of Hotel Customer Reviews for Sustainable Tourism."
- Mola, Sebastianus Adi Santoso, Rosa V. K. Isabela Ohe Roma, and Tiwuk Widiastuti. 2025. *Text Mining:*

Analisis Sentimen Dengan Lexicon. Kaizen Media Publishing.

Rahman, Mostafizer, Ariful Islam Shiplu, Yutaka Watanobe, and Ashad Alam. 2021. "RoBERTa-BiLSTM : A Context-Aware Hybrid Model for Sentiment Analysis." 14(8):1–18.

Rahmatullah, Tansah. 2021. "Teknologi Persuasif: Aktor Penting Media Sosial Dalam Mengubah Sikap Dan Perilaku Pengguna." *Jurnal Sosbum Insentif* 4(1):60–78.

Semary, Noura A., Wesam Ahmed, Khalid Amin, and Pawel Plawiak Id. 2024. "Enhancing Machine Learning-Based Sentiment Analysis through Feature Extraction Techniques." doi:10.1371/journal.pone.0294968.

Setiadarma, Aan, Ahmad Zaki Abdullah, Priyono Sadjijo, and Dwi Firmansyah. 2024. "Tinjauan Literatur Transformasi Sosial Dalam Era Virtual." *Khatulistiwa: Jurnal Pendidikan Dan Sosial Humaniora* 4(1):232–44.

Srg, Rika Aulia Maharani, and Usiono Usiono. 2024. "Peran Media Digital Dalam Meningkatkan Keterlibatan Dan Partisipasi Publik: Transformasi Komunikasi Di Era Informasi Dan Sosial." *Journal Sains Student Research* 2(6):506–13.

Sulaiman, Agus, Tri Basuki Kurniawan, Deshinta Arrova Dewi, and Mashal Alqudah. 2025. "Utilizing Sentiment Analysis for Reflect and Improve Education in Indonesia." 6(1):189–200.

Widagdo, Adika Sri, and Fachruddin Edi Nugroho Saputro. 2024. *Buku Monograf Analisis Sentimen Mobil Listrik Berbasis Ulasan Youtube*. PT. Sonpedia Publishing Indonesia.

Internet

CNN Indonesia. 2026. "FULL Momen Peserta Cerdas Cermat Empat Pilar MPR RI Protes Ke Juri." <https://www.youtube.com/watch?v=m1NJSX-NTCw>.