Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2629~2638
e-ISSN: 2548-6861

Performance Analysis of BERT and CLIP Models in Multimodal Sentiment Classification of Short Video Content

Very Dwi Setiawan[1]*, Endang Anggiratih[2]*, Najwa Eka Putriningsih[3]*, Jonathan Eldo Kusuma[4]*
* Department of Informatics, Faculty of Science and Technology, Pignatelli Triputra University, Indonesia
ferystiawan54@gmail.com [1], enratih86@gmail.com[2], najwaekap034@gmail.com[3], eldokusuma3@gmail.com[4]

Article Info

Article history:

Received 2026-04-21
Revised 2026-05-25
Accepted 2026-06-11

Keyword:

CLIP,
IndoBERT,
Multimodal,
Sentiment Analysist,
Short Video,
Transformer Model.

ABSTRACT

The rapid growth of short video platforms such as YouTube Shorts has increased the need for effective sentiment analysis methods capable of capturing public opinion in multimodal content. This study analyzes and compares the effectiveness of unimodal and multimodal approaches for sentiment classification of Indonesian short videos, focusing on IndoBERT for text-based modeling and CLIP for multimodal integration. The main objective is to investigate whether incorporating visual information alongside textual data can improve sentiment classification performance compared to a text-only approach. The dataset consists of 1,128 Indonesian short videos collected from YouTube Shorts. Audio data are transcribed into text using Automatic Speech Recognition (ASR), while visual information is represented using video thumbnails. Sentiment labels are automatically categorized into three classes (positive, neutral, and negative) using a pre-trained IndoBERT model. In the training phase, the unimodal approach relies solely on textual features extracted by IndoBERT, whereas the multimodal approach integrates textual and visual features using CLIP through feature-level fusion. Model performance is evaluated using accuracy, precision, recall, F1-score, and computational time analysis. The experimental results show that the unimodal text-based model outperforms the multimodal model, achieving higher accuracy (86% vs 82%) and better overall evaluation metrics. IndoBERT also demonstrates better convergence behavior compared to English BERT, with training accuracy increasing from 0.76 to 0.86 and validation accuracy from 0.77 to 0.88, along with lower loss values. In contrast, English BERT achieves lower performance, with training accuracy rising from 0.72 to 0.79 and validation accuracy from 0.73 to 0.80. Furthermore, the unimodal approach requires significantly less computation time (18 minutes compared to 35 minutes). These findings indicate that textual information plays a dominant role in sentiment expression in Indonesian short video content, while visual features increase computational complexity without significant performance gains.

I. INTRODUCTION

The rapid advancement of digital technology has driven a major transformation in the way people consume and share information [1]. Social media applications such as TikTok, Instagram Reels, and YouTube Shorts have widespread brief type video content, which has now become one of the primary media for expressing opinions, product reviews, and social campaigns. Short videos do not only contain text-based information (captions, subtitles), but also incorporate visual elements (facial expressions, gestures, objects) and audio components (voice intonation, music). This content complexity presents significant opportunities for richer sentiment analysis, while simultaneously introducing new

challenges in capturing and understanding emotions embedded within the combination of these modalities [2],[3].

Natural Language Processing (NLP) is a domain of artificial intelligence that concentrates regarding comprehension and handling human language in both textual and spoken forms [4], [5]. One of its prominent applications is sentiment analysis, which aims to identify, extract, and classify opinions into positive, negative, or neutral categories. Sentiment analysis research has traditionally relied heavily on written data, like as consumer feedback or social media comments [6]. While this approach has proven effective in many scenarios, it becomes limited when analyzing content that involves strong visual and audio contexts. For instance, a joyful facial expression accompanied by sarcastic text may lead to different sentiment interpretations when only a single modality is considered [7]. To address these limitations, Multimodal Sentiment Analysis (MSA) has appeared as an approach that combines text, images, or audio to provide a greater thorough comprehension of sentiment [8].

In the design of MSA, deep learning-based architectures serve an essential function [9]. Bidirectional Encoder Representations from Transformers (BERT), a transformer-based architecture developed bidirectionally to represent situational associations between words in a sentence, has become one of the dominant models in text processing due to its strong contextual representation capabilities [10]. Consequently, BERT is frequently used as the primary component for textual feature extraction in MSA tasks. On the other hand, Contrastive Language Image Pretraining (CLIP), a architectures developed by OpenAI that employs contrastive learning to align textual and visual representations within a shared embedding space, is well suited for analyzing visual-textual content [11]. CLIP has been widely adopted in MSA to capture semantic relationships between text and visual elements [12]. Although both models have demonstrated strong performance within their respective domains, prior studies have largely focused on global languages such as English. In contrast, comparative studies examining the effectiveness of BERT and CLIP specifically in the context of multimodal sentiment analysis for Indonesian short videos remain limited, despite the unique linguistic characteristics of the Indonesian language [13].

Several previous studies have explored multimodal sentiment analysis using various approaches. Zhao et al. [14] proposed the HIMT and JAAIN models for multimodal sentiment analysis, achieving an F1-score of 80% with improvements of approximately 3 until 10 percentage points, demonstrating more accurate and consistent performance across datasets such as Twitter-15, Twitter-17, and multi-ZOL. Qi et al. [27] introduced the Multimodal Encoding Decoding System with Transformer (MEDT), which showed improved classification performance, attaining an accuracy of 80% and F1-score of 85%. Mu et al. [15] employed a multimodal framework based on BERT, CLIP, LoRA, and cross-attention fusion, attaining an accuracy of 96.80% and F1-score of 94.23%, representing an improvement of more than 11% compared to unimodal models. Sehar et al. [16] utilized a multimodal architecture combining LSTM for text and CNN for images, achieving accuracy of 87% and F1-score of 85%, outperforming traditional methods by approximately 8 until 10%.

Although multimodal sentiment analysis has been widely explored, most previous studies focus on English-language datasets and employ computationally intensive architectures using full-video or frame-sequence representations. Studies comparing lightweight unimodal and multimodal approaches for Indonesian short-video sentiment analysis remain limited, particularly in terms of computational efficiency and the trade-off between model complexity and classification performance on platforms such as YouTube Shorts.

To address this gap, this study compares a text-based unimodal approach using IndoBERT with a lightweight multimodal approach integrating textual and visual information using CLIP. Unlike prior studies that rely on dense temporal video representations, this research utilizes video thumbnails as lightweight visual features to reduce computational complexity while preserving essential visual context. The main contributions of this study include evaluating unimodal and multimodal sentiment classification on Indonesian YouTube Shorts datasets, analyzing performance and computation-time trade-offs, and highlighting the practical limitations of lightweight multimodal representations for short-video sentiment analysis.

II. METHODOLOGY

This section describes the stages and processing techniques employed in this study, beginning with the collection of a dataset consisting of short videos. Each video is subsequently decomposed into multiple modalities, including audio data that are converted into text using Automatic Speech Recognition (ASR), as well as visual components extracted in the form of representative clips or structures. According to this step, preprocessing is utilized to both textual and visual data to remove noise that may interfere with the analysis process. The processed data are then assigned sentiment labels and transformed into numerical representations through feature extraction, utilizing the BERT model for the textual modality and the CLIP model for the visual modality. The final stage of the study involves model training and performance evaluation under unimodal and multimodal scenarios, followed by a comparative analysis of the effectiveness of the two approaches. The overall research workflow is illustrated in Figure 1.

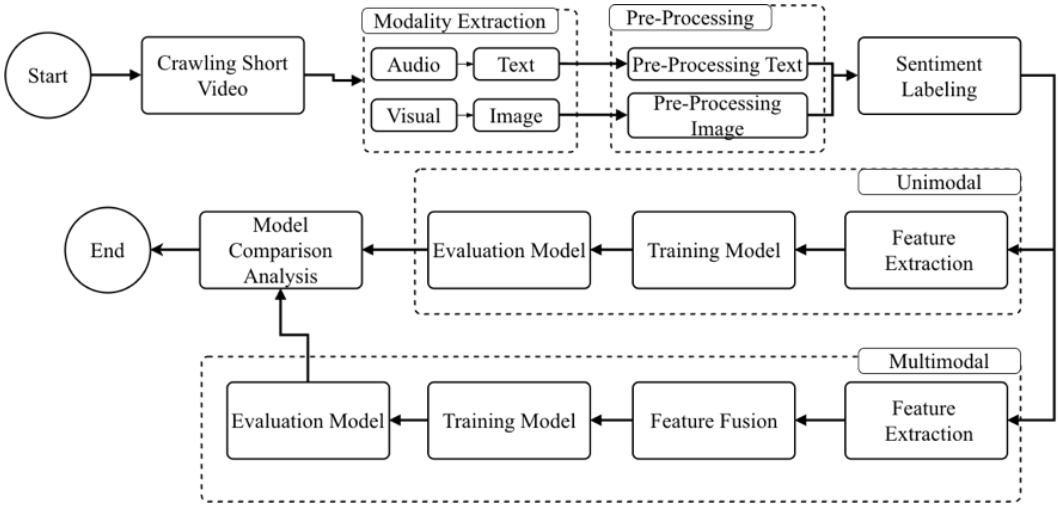


Figure 1. Research flow

A. *Dataset*

The short video data crawling process was conducted using the YouTube Data API to collect Indonesian-language YouTube Shorts uploaded between January 2025 and December 2025, resulting in a total of 1,128 videos. The crawling procedure began with the definition of search keywords and the application of a duration filter to restrict the results to short videos with a maximum length of 60 seconds. The system then automatically retrieved relevant video information, including video ID, title, description/caption, upload date, duration, and video URL. Subsequently, the video files were downloaded for the purposes of audio and visual extraction, while the associated metadata were stored to support further analysis. The resulting dataset was further filtered to ensure adequate audio and visual quality as well as conformity with the Indonesian language. The final dataset obtained after the download process is presented in Table 1.

TABLE I
DATASET SAMPLE

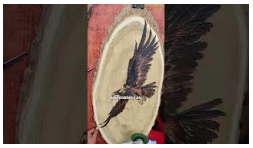
ID	Title	Duration
1ykFZKe4UI	Musibah menciptakan ikatan diantara 2 Klub ini	38 detik
hPTjvAsdNWY	Habis lari nyobain makanan khusus TNI| Rasanya ada yg tau???	02:25 Menit
AKx2qdk43-U	There's Nothing Wrong with Vino Pre	14 Detik
fQja2sy2ooQ	Berawal dari gerinda menjadi sebuah karya seni yang indah	10 Detik

B. *Modality Extraction*

The modality extraction stage in this study begins with the primary data in the form of short videos collected through the crawling process. In the initial step, each video is processed to separate the audio and visual components as distinct sources of sentiment information. For the audio modality, the sound track is extracted from the video using audio extraction techniques, resulting in audio files in a standard format. The extracted audio is then transcribed into text using ASR technology, which converts spoken language in the video into a textual representation that can be further processed. This transcription is regarded as a verbal-based sentiment representation that reflects the speaker's opinions, emotions, or expressions conveyed in the video. For the visual modality, this study adopts a simpler and more efficient approach by extracting only the thumbnail of each video as the primary visual representation. Thumbnails are selected because they are generally designed to succinctly and attractively represent the core content of a video, thereby containing essential visual information that may reflect sentiment, such as facial expressions, dominant objects, or specific visual contexts. The results of the audio-to-text extraction process are presented in Table 2.

TABLE 2
AUDIO TO TEXT EXTRACTION MODALITIES

Video Id	Title	Transcript	Thumbnail
1ykFZKe4UI	Musibah menciptakan ikatan diantara 2 Klub ini #sepakbola #ferryirwandi #coachjustin	ternyata gue baru tahu juga bahwa Benfica itu menganggap Torino itu sister klubnya karena kejadian pesawat pertandingan di Benfica dia ternyata gitu Yang banyak berangkat atau pulang pulang dari Benfica yang pemain Timnas Italy semuanya udah berapa orang di situ ada 8 orang kalau nggak salah meninggal yang kemudian Torino harus tidak pernah	

Video Id	Title	Transcript	Thumbnail
		kembali lagi ke orbitnya yang itu kan karena kalau itu tidak terjadi peta kekuasaan vitalitas	
hPTjvAsdNWY	Habis lari nyobain makanan khusus TNI| Rasanya ada yg tau???	jadi para-para TNI bawa ini semua bawa kalau habis baru tembak babi dan abang ini muslim muslim makanan makanannya nggak boleh diperjual belikan juga emang itu makanan tentang seumur hidup nggak pernah makan makanan di atas sini bolak-balik di bawah itu diatas termasuk TNI teman-teman ini pelajaran konten ini teman-teman oh ini sumur itu belum pernah nyobain apa Nih Hmm kewajibanmu kaya rempah-rempah nasinya mungkin gara-gara dipanasin nya jadi nasi lama ya agak lembek dikit mau mau bubur tapi bukan bubur dan dagingnya enak kalau bertahan di hutan makan ini sih kenyang sih tapi bakal lapar lagi soalnya kan tentara harus segera terus ya jadi para para TNI kalau ke hutan bawa ini semua bawa kalau habis tembak babi Jadi Abang pernah makan babi Enggak apa-apa ini jadinya kita di hutan kan aku pernah di hutan 3 bulan biar kehabisan makanan helikopter juga belum sampai pas di Papua itu katanya lewat kalau kita tembak burung hantu dan posisinya bergerak terus gimana menurut teman-teman yang pertama tapi kalau ada makanan lain kita makan babi iya berdosa di mana rasanya babi mohon maaf ya untuk istilah bukannya apa ya enak tapi cukup untuk mengisi Perut saya sakit memang babi teman-teman	
fQja2sy2ooQ	Berawal dari gerinda menjadi sebuah karya seni yang indah	Baru tahu ya kalau alat potong besi yaitu gerinda bisa dipakai melukis di atas papan kayu caranya dengan menggesekkan mata gerinda di atas papan kayu tersebut sehingga menimbulkan warna coklat kehitaman keren banget	

C. Preprocessing

The text preprocessing stage in this study aims to clean and normalize textual data obtained from ASR transcriptions and video descriptions so that it is suitable for characteristic acquisition and architecture training [17]. The process begins with text denoising, which includes the removal of HTML tags using strip_html, elimination of text enclosed in square brackets through remove_between_square_brackets, and normalization of informal words or abbreviations by converting contractions into their full forms using replace_contractions. Subsequently, the text is очищed from technical artifacts such as byte string prefixes (remove_byte_str), retweet and mention markers (remove_retweet_mention), and social media usernames (remove_username). To ensure character consistency, non-ASCII characters are normalized into standard ASCII using remove_nonascii [18]. The next step removes elements that are irrelevant to sentiment analysis, including URLs (remove_url), numerical digits (remove_digit), and punctuation marks (remove_punctuations). In addition, single letters that do not carry semantic meaning (remove_single_letter) are also eliminated. Excessive whitespace is normalized using remove_additional_white_spaces, and aims text is changed to lowercase (lowercasing) to avoid semantic inconsistencies caused by capitalization. This preprocessing pipeline produces cleaner, more consistent, and more informative text, thereby enhancing the quality of feature representations and improving the performance of the sentiment analysis model. The outcomes of the preprocessing stage are illustrated in Table 3.

TABLE 3
PREPROCESSING.

Video Id	Transcript	Text Clean
1ykFZKe4UI	ternyata gue baru tahu juga bahwa Benfica itu menganggap Torino itu sister klubnya karena kejadian pesawat pertandingan di Benfica dia ternyata gitu Yang banyak berangkat atau pulang pulang dari Benfica yang pemain Timnas Italy semuanya udah berapa orang di situ ada 8 orang kalau nggak salah meninggal yang kemudian Torino harus tidak pernah kembali lagi ke orbitnya yang itu kan karena kalau itu tidak terjadi peta kekuasaan vitalitas	ternyata saya baru mengetahui bahwa benfica menganggap torino sebagai klub saudara karena peristiwa kecelakaan pesawat yang melibatkan tim benfica. banyak pemain tim nasional italia yang berangkat atau pulang dari benfica dan meninggal dunia dalam peristiwa tersebut. akibat kejadian itu, torino tidak pernah kembali ke masa kejayaannya karena perubahan peta kekuatan sepak bola
hPTjvAsdNWY	jadi para-para TNI bawa ini semua bawa kalau habis baru tembak babi dan abang ini muslim muslim makanan makanannya nggak boleh diperjual belikan juga emang itu makanan tentang seumur hidup nggak pernah makan makanan di atas sini bolak-balik di bawah itu diatas termasuk TNI teman-teman ini pelajaran konten ini teman-teman oh ini sumur itu belum pernah nyobain apa Nih Hmm kewajibanmu kaya rempah-rempah nasinya mungkin gara-gara dipanasin nya jadi nasi lama ya agak lembek dikit mau mau bubur tapi bukan bubur dan dagingnya enak kalau bertahan di hutan makan ini sih kenyang sih tapi bakal lapar lagi soalnya kan tentara harus segera terus ya jadi para para TNI kalau ke hutan bawa ini semua bawa kalau habis tembak babi Jadi Abang pernah makan babi Enggak apa-apa ini jadinya kita di hutan kan aku pernah di hutan 3 bulan biar kehabisan makanan helikopter juga belum sampai pas di Papua itu katanya lewat kalau kita tembak burung hantu dan posisinya bergerak terus gimana menurut teman-teman yang pertama tapi kalau ada makanan lain kita makan babi iya berdosa di mana rasanya babi mohon maaf ya untuk istilah bukannya apa ya enak tapi cukup untuk mengisi Perut saya sakit memang babi teman-teman	para anggota tni bawa seluruh bekal jika sedia habis mereka buru babi di hutan narasumber agama islam sehingga makan tidak boleh jual dan tidak pernah konsumsi sepanjang hidup konten beri ajar bahwa makan itu hanya konsumsi dalam kondisi darurat nasi rasa lembek karena panas ulang hampir serupa bubur namun daging rasa cukup enak dan kenyang saat tugas di hutan makan ini cukup tahan hidup meski lapar muncul kembali karena aktivitas fisik berat narasumber cerita alam tugas di papua selama tiga bulan tanpa pasok makan hingga helikopter bantu tiba dalam kondisi tentu konsumsi makan itu laku karena paksa bukan pilih meski secara agama anggap dosa
fQja2sy2ooQ	Baru tahu ya kalau alat potong besi yaitu gerinda bisa dipakai melukis di atas papan kayu caranya dengan menggesekkan mata gerinda di atas papan kayu tersebut sehingga menimbulkan warna coklat kehitaman keren banget	baru tahu alat potong besi berupa gerinda dapat guna lukis atas papan kayu teknik laku cara gesek mata gerinda pada muka kayu hingga hasil warna coklat hitam yang lihat sangat tarik

D. Labeling

The sentiment labeling process in this study is conducted automatically using a lexicon-based sentiment approach integrated with the pretrained Transformer model mdhugol/indonesia-bert-sentiment-classification. The lexicon-based technique utilizes a predefined Indonesian sentiment dictionary containing positive, negative, and neutral vocabularies that have been previously validated in sentiment analysis research, allowing each word or expression in the textual content to be associated with sentiment polarity scores based on established linguistic sentiment representations. This approach helps reduce uncertainty in sentiment assignment while improving the reliability of the generated labels. Furthermore, the pretrained BERT-based model is employed to capture contextual semantic relationships and emotional nuances in Indonesian text, including both formal and informal expressions commonly found in short-video content. The preprocessed textual data obtained from ASR transcriptions and video captions are processed using the HuggingFace Transformers pipeline to generate contextual representations and sentiment probabilities for three categories: positive, neutral, and negative. The final sentiment label is determined based on the dominant sentiment score and contextual probability estimation. This combined strategy is selected to improve

labeling consistency, scalability, and contextual understanding while maintaining sentiment reliability through lexicon-based validation.

E. Training Model

In the first stage, this study adopts a unimodal approach by utilizing IndoBERT as the primary model for text-based sentiment analysis [19]. The textual data are obtained from video captions and audio transcriptions (ASR) that have undergone preprocessing and sentiment labeling. IndoBERT is employed to extract contextual and bidirectional textual feature representations, enabling the model to capture semantic meaning, linguistic structure, and emotional nuances in the Indonesian language[20]. The resulting vector representations are then passed to a classification head to predict sentiment classes, namely positive, neutral, and negative. This unimodal approach serves as a baseline to evaluate sentiment analysis performance using textual information alone, without incorporating additional modalities. The BERT architecture used in this study is illustrated in Figure 2.

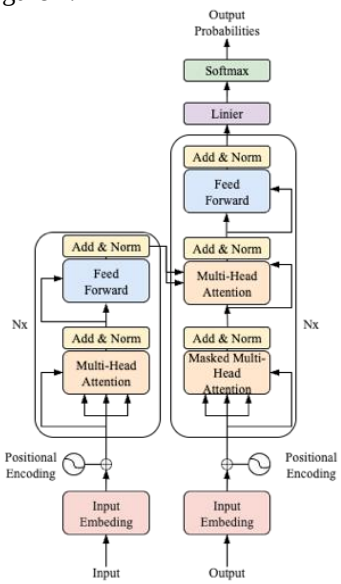


Figure 2. BERT Architecture [21].

In the second stage, this study develops a multimodal approach by integrating textual and visual information using the CLIP model. The textual modality is processed through the CLIP text encoder, while the visual modality is obtained from video thumbnails and processed using the CLIP image encoder. Both encoders generate feature representations within a shared embedding space, enabling effective cross-modal integration. Subsequently, feature-level fusion is performed by combining the textual and visual vectors into a single multimodal representation. This combined encoding is then utilized as features for the sentiment classification model [22]. The multimodal approach attempts to strengthen sentiment analysis performance by incorporating visual context that cannot be fully captured through textual information alone. This study intentionally uses thumbnails as lightweight visual representations to reduce computational complexity, although this may limit temporal and emotional context extraction. The CLIP architecture utilized in this study is illustrated in Figure 3.

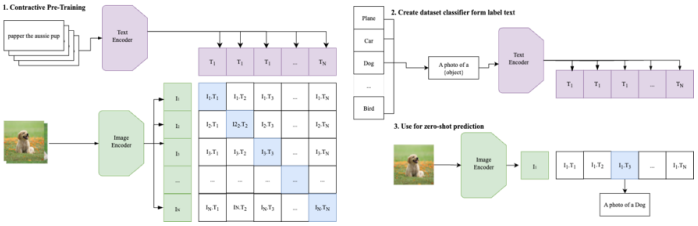


Figure 3. CLIP Architecture [23].

Pada model CLIP, feature-level fusion dilakukan dengan terlebih dahulu mengekstraksi fitur dari dua modalitas, yaitu gambar dan teks, menggunakan image encoder dan text encoder. Masing-masing encoder menghasilkan representasi fitur berupa embedding vektor berdimensi tertentu yang menangkap informasi semantik dari masing-masing modalitas. Selanjutnya, kedua vektor tersebut digabungkan pada level fitur (feature level) menggunakan metode seperti concatenation sehingga membentuk satu vektor representasi gabungan. Vektor hasil penggabungan ini kemudian digunakan sebagai input pada tahap selanjutnya, seperti klasifikasi atau pengukuran kesamaan, sehingga model dapat memanfaatkan informasi dari kedua modalitas secara bersamaan.

F. Model Evaluation

Model evaluation was performed using test data to measure accuracy, precision, recall, and F1-score [24]. Additionally, a confusion matrix was used to detail correct and incorrect predictions for each sentiment class, helping identify areas for model improvement.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recal}} \quad (4)$$

True Positive (TP) indicates to the count of examples which are estimated as positive by the model and are indeed correctly classified as belonging to the positive category. True Negative (TN) represents the count of examples that are estimated as negative and correctly correspond to the actual negative category. False Positive (FP) denotes cases that are predicted as positive by the model but actually belong to the negative classification, whereas False Negative (FN) corresponds to occurrences that are predicted as negative even though they truly belong to the positive class

III. Result and Discussion

This section presents the experimental results, including the sentiment labeling process as well as the training outcomes of the unimodal and multimodal models. The results of the data preprocessing stage have been described in detail in the methodology section. Sentiment labeling in this study was conducted on a dataset consisting of 1,128 short videos with three sentiment label, positive, negative, and neutral, using IndoBERT as the primary model for automatic labeling prior to the model training phase. Out of the total 1,128 data instances, the sentiment distribution comprises 462 positive samples 41.0%, 389 neutral samples 34.5%, and 277 negative samples 24.5%. This distribution indicates a dominance of positive and neutral sentiment content with a relatively balanced class proportion, which supports a fairer training and evaluation process when comparing multimodal and non-multimodal models. The distribution of the sentiment labels is illustrated in Figure 4.

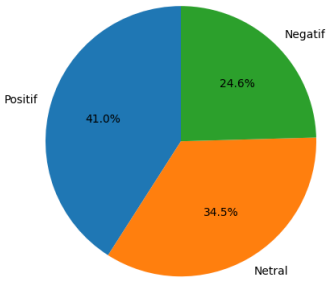


Figure 4. Distribution Data Sentiment.

Based on Figure 4, the sentiment labeling results indicate that the class distribution is relatively balanced; therefore, no data augmentation or class balancing techniques are required. The dataset used in this study consists of Indonesian language YouTube short videos covering a wide range of topics and representing diverse perspectives. The textual content extracted from these short videos is subsequently visualized in the form of a word cloud to illustrate the dominant words appearing in the dataset, as shown in Figure 5.

Figure 5. Word Cloud.

The ASR quality evaluation shown in Figure 6 was measured using Word Error Rate (WER) and Character Error Rate (CER). The obtained results show a WER value of 0.18 and a CER value of 0.09, indicating that the transcription process produced an error rate of 18% at the word level and 9% at the character level. These results demonstrate that the ASR system is capable of generating relatively accurate transcriptions for sentiment analysis tasks. Lower WER and CER values indicate better transcription quality, which contributes to more reliable textual representations for the sentiment classification process.

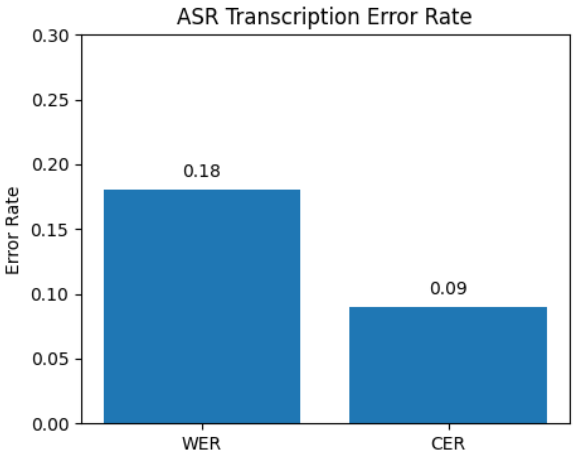


Figure 6. ASR Transcription Eror Rate.

The architecture training in this study employs IndoBERT for both unimodal and multimodal approaches, using evaluation measures incorporating accuracy, precision, recall, and F1-score. The outcomes of the architecture training are presented in Figure 7.

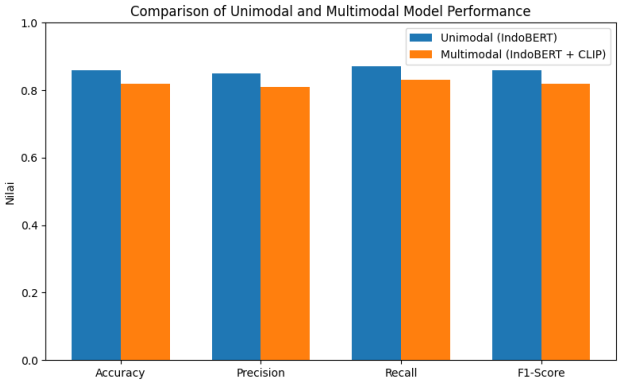


Figure 7. Comparison of Unimodal and Multimodal Model Performance.

The assessment outcomes presented in Figure 6 demonstrate which the unimodal model outperforms the multimodal model. The unimodal model achieved an accuracy of 0.86, precision of 0.85, recall of 0.87, and F1-score of 0.86, whereas the multimodal model obtained an accuracy of 0.82, precision of 0.81, recall of 0.83, and F1-score of 0.82. As a comparison, this study uses an English-based BERT model which can be seen in Figure 8.

In Figure 8, the comparison below shows that the IndoBERT model performs better than BERT English in Indonesian sentiment analysis. IndoBERT achieved an accuracy of 0.79 and an F1-score of 0.77, while BERT English only achieved an accuracy of 0.74 and an F1-score of 0.74. This is because IndoBERT was trained on an Indonesian corpus, making it better able to understand vocabulary.

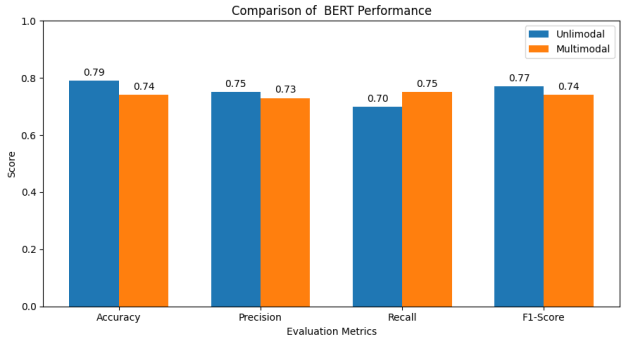


Figure 8. Comparison of Unimodal and Multimodal Using BERT Model.

The BERT English model is more optimized for English, making it less effective at capturing semantic patterns and sentiment in Indonesian.

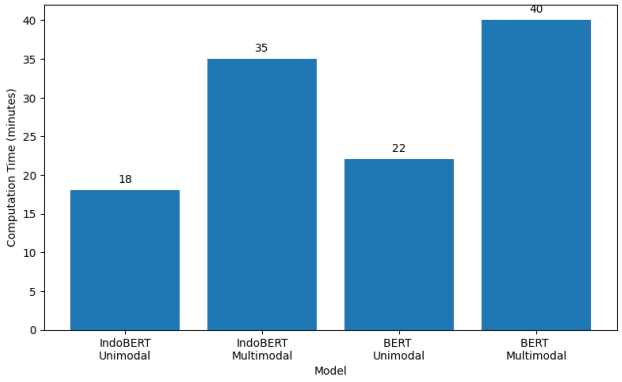


Figure 9. Comparison Time Computations.

The computation time comparison shows figure 9 that the IndoBERT Unimodal model achieves the fastest processing time compared to the other architectures, requiring approximately 18 minutes, while IndoBERT Multimodal requires around 35 minutes due to the additional complexity of combining textual and visual features. Meanwhile, the BERT English Unimodal model requires approximately 22 minutes, and the BERT English Multimodal model records the highest computation time at around 40 minutes. These results indicate that multimodal architectures significantly increase computational cost because the model must process multiple modalities simultaneously, including feature extraction, fusion mechanisms, and optimization of textual and visual representations.

In addition, IndoBERT demonstrates better computational efficiency than the English-based BERT model because it is specifically pre-trained on Indonesian corpora, enabling faster convergence and more effective semantic understanding for Indonesian sentiment analysis tasks. All experiments were conducted using Google Colab with NVIDIA Tesla T4 GPU acceleration and 32 GB RAM, where the Tesla T4 GPU supports faster transformer-based deep learning computations through CUDA acceleration, while the 32 GB RAM enables efficient dataset handling, multimodal feature extraction, and batch processing during training and evaluation.

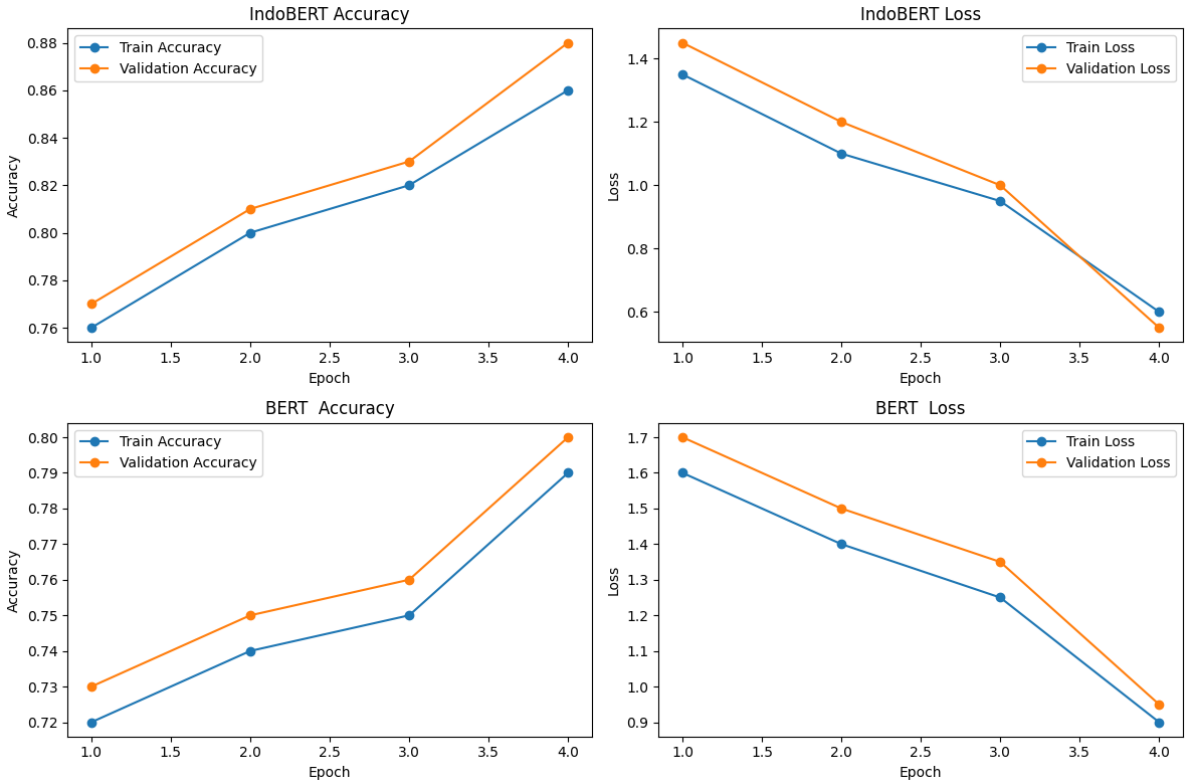


Figure 10. Training accuracy per epoch for Unimodal and Multimodal.

Figure 10 shows that IndoBERT achieves more stable and higher performance than the English BERT model across training epochs. IndoBERT reaches a validation accuracy of 0.88 and a lower validation loss of 0.55, indicating better convergence and learning efficiency, while BERT English only achieves 0.80 accuracy with higher loss. This difference occurs because IndoBERT is trained on Indonesian corpora, enabling better understanding of local language structures, slang, and contextual expressions. Consequently, IndoBERT demonstrates superior contextual and semantic representation for Indonesian sentiment analysis compared to BERT English.

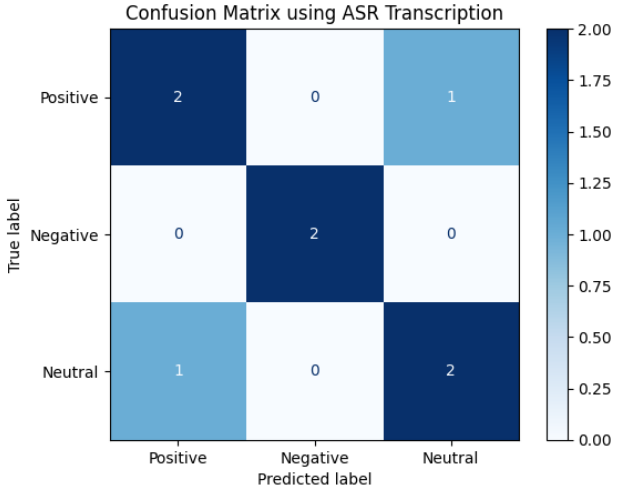


Figure 11. Consufion Matrix using ASR Transcirption.

The confusion matrix shown in Figure 11 shows that the model obtained 6 correct predictions from 8 test datasets with an accuracy of around 0.75, with the Negative class performing best with no misclassifications. Meanwhile, errors occurred in the Positive and Neutral classes, which were swapped by one dataset each, indicating that the model still had difficulty distinguishing similar sentiment expressions. This indicates that noise from the ASR transcription results can affect the model's accuracy in understanding the overall sentiment context.

The experimental results demonstrate that the IndoBERT-based unimodal model achieves better performance compared to the multimodal architecture in this study of Indonesian short-video sentiment analysis. The unimodal model obtained higher accuracy, precision, recall, and F1-score than the multimodal model under the current experimental configuration. This finding suggests that textual information extracted from ASR transcripts provides stronger semantic and emotional cues for sentiment classification within the dataset used in this research. By utilizing IndoBERT, which is specifically pre-trained on large-scale Indonesian corpora, the unimodal approach can generate richer semantic and contextual representations for Indonesian-language content. In contrast, the multimodal model combines textual and visual features, resulting in a more complex architecture and higher computational requirements. Although multimodal learning theoretically provides complementary information from multiple modalities, the integration process introduces challenges in feature fusion and optimization. In this study, the visual modality was represented only by video thumbnails, which may not fully capture the temporal, emotional, and contextual information contained throughout the videos. Consequently, the visual representation may introduce less relevant information during training, reducing the effectiveness of sentiment classification and slowing the convergence process compared to the unimodal approach.

Furthermore, the comparison between IndoBERT and the English-based BERT model shows that IndoBERT consistently outperforms BERT English in both unimodal and multimodal experiments. IndoBERT achieves higher validation accuracy and lower loss values because it is specifically optimized for Indonesian linguistic characteristics, including informal vocabulary, abbreviations, local expressions, and contextual semantics commonly found in Indonesian short-video content. Meanwhile, the English BERT model is primarily trained on English corpora, limiting its ability to accurately capture Indonesian semantic structures and contextual sentiment information. Overall, the experimental findings indicate that the IndoBERT unimodal model provides the best balance between performance, efficiency, and computational cost within the current experimental setup, while richer visual representations may further improve multimodal performance in future studies. Although the proposed approach demonstrates promising performance on Indonesian YouTube Shorts datasets, the generalizability of the findings to other short-video platforms remain an important consideration. Different platforms such as TikTok, Instagram Reels, and Facebook Reels may exhibit distinct characteristics in terms of video editing styles, user interaction patterns, audio quality, language usage, and visual presentation. These variations may influence the effectiveness of both unimodal and multimodal sentiment analysis models. Therefore, future research should evaluate the proposed approach on cross-platform datasets to further investigate the robustness and adaptability of the model across diverse short-video environments.

IV. CONCLUSION

Based on the results of this study, the text-based unimodal approach using IndoBERT demonstrates better performance than the multimodal approach for Indonesian short-video sentiment analysis. The unimodal model achieves higher accuracy, precision, recall, and F1-score, while also requiring shorter computation time. These findings indicate that textual information provides more dominant and relevant sentiment representations compared to visual information in short-video content. On the other hand, the multimodal approach introduces higher architectural complexity and longer computation time without significant performance improvement. Therefore, future research is recommended to develop more effective multimodal fusion strategies, utilize

visual features that are more semantically related to sentiment, and evaluate the proposed approach using larger datasets from multiple short-video platforms such as TikTok and Instagram Reels to improve robustness and generalization capability.

REFERENCES

[1] Z. Van Veldhoven and J. Vanthienen, "Digital transformation as an interaction-driven perspective between business, society, and technology," *Electron. Mark.*, vol. 32, no. 2, pp. 629–644, 2022, doi: 10.1007/s12525-021-00464-5.

[2] L. Theodorakopoulos, A. Theodoropoulou, and C. Klavdianos, "Interactive Viral Marketing Through Big Data Analytics, Influencer Networks, AI Integration, and Ethical Dimensions," *J. Theor. Appl. Electron. Commer. Res.*, vol. 20, no. 2, 2025, doi: 10.3390/jtaer20020115.

[3] S. Nur, S. Sahibu, and M. Razak, "Aspect-Based Sentiment Analysis of Tourist Attractions in Labuanbajo Using the Transformer Model as a Recommendation for Improving Service Quality," vol. 10, no. 1, pp. 496–502, 2026.

[4] E. G. Prasetio, L. B. Handoko, and K. Hastuti, "Improving Retrieval-Augmented Generation Performance Using the MAF-RAG Architecture , EVR – VOR Vector Retrieval , and Multi-Agent Fallback Reasoning," vol. 10, no. 1, pp. 212–223, 2026.

[5] V. D. Setiawan and D. U. Iswavigra, "Sentiment Analysis to Evaluate Public Service Perception among Surakarta City Residents Using the BiLSTM Model," *J. Informatics Telecommun. Eng.*, vol. 9, no. July, pp. 229–239, 2025.

[6] F. Santosa, E. Oktafanda, H. Setiawan, and A. Latif, "Advanced Long Short-Term Memory (LSTM) Models for Forecasting Indonesian Stock Prices," *J. Galaksi*, vol. 1, no. 3, pp. 198–208, 2024, doi: 10.70103/galaksi.v1i3.42.

[7] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Nat. Lang. Process. J.*, vol. 6, no. February, p. 100059, 2024, doi: 10.1016/j.nlp.2024.100059.

[8] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Inf. Fusion*, vol. 101, p. 101973, 2024, doi: https://doi.org/10.1016/j.inffus.2023.101973.

[9] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "{M}-{SENA}: An Integrated Platform for Multimodal Sentiment Analysis," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, May 2022, pp. 204–213, doi: 10.18653/v1/2022.acl-demo.20.

[10] Y. Sun, H. Yuan, and F. Xu, "Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features," *Nat. Lang. Process. J.*, vol. 11, p. 100148, 2025, doi: https://doi.org/10.1016/j.nlp.2025.100148.

[11] A. D. Dobrzycki, A. M. Bernardos, L. Bergesio, A. Pomirski, and D. Sáez-Trigueros, "Exploring the Use of Contrastive Language-Image Pre-Training for Human Posture Classification: Insights from Yoga Pose Analysis," *Mathematics*, vol. 12, no. 1, 2024, doi: 10.3390/math12010076.

[12] X. Pan, T. Ye, D. Han, S. Song, and G. Huang, "Contrastive Language-Image Pre-Training with Knowledge Graphs," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.

[13] Z. Lu, H. Li, N. A. Parikh, J. R. Dillman, and L. He, "RadCLIP: Enhancing Radiologic Image Analysis Through Contrastive Language–Image Pretraining," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 10, pp. 17613–17622, 2025, doi: 10.1109/TNNLS.2025.3568036.

[14] H. Zhao, M. Yang, X. Bai, and H. Liu, "A Survey on Multimodal Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 12, pp. 12039–12052, 2024, doi: 10.1109/ACCESS.2024.3354844.

[15] G. Mu, C. Chen, X. Li, J. Li, X. Ju, and J. Dai, "Multimodal Sentiment Analysis of Government Information Comments Based on Contrastive Learning and Cross-Attention Fusion Networks," *IEEE Access*, vol. 12, no. November, pp. 165525–165538, 2024, doi: 10.1109/ACCESS.2024.3493933.

[16] U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, and F. Khan, "Urdu Sentiment Analysis via Multimodal Data Mining Based on Deep Learning Algorithms," *IEEE Access*, vol. 9, pp. 153072–153082, 2021, doi: 10.1109/ACCESS.2021.3122025.

[17] M. R. Nursyam, M. Koprawi, and D. Ariyus, "Optimizing Email Spam Detection through Handling Class Imbalance with Class Weights and Hyperparameter Using GridSearchCV," vol. 10, no. 1, pp. 232–244, 2026.

[18] I. Muslim, M. Firdaus, and R. Habibi, "Named Entity Recognition in Indonesian History Textbook Using BERT Model," *CogITo Smart J.*, vol. 11, no. 1, pp. 140–151, 2025, doi: 10.31154/cogito.v11i1.880.140-151.

[19] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, "BERT base model for toxic comment analysis on Indonesian social media," *Procedia Comput. Sci.*, vol. 216, pp. 714–721, 2023, doi: https://doi.org/10.1016/j.procs.2022.12.188.

[20] D. U. Iswavigra, V. D. Setiawan, M. Ulfa, and B. Ommr, "Sentiment Analysis Using Bidirectional Encoder Representations from Transformers for Indonesian Stock Price Prediction with Long Short-Term Memory and Gated Recurrent Unit Models," vol. 7, no. 2, pp. 961–976, 2026.

[21] V. D. Setiawan, D. U. Iswavigra, and E. Anggiratih, "Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court ' s Decision Regarding the Minimum Age of Vice Presidential Candidates," *Sci. J. Informatics*, vol. 12, no. 3, pp. 397–406, 2025, doi: 10.15294/sji.v12i3.26360.

[22] H. You *et al.*, "Learning Visual Representation from Modality-Shared Contrastive Language-Image Pre-training," in *Computer Vision -- ECCV 2022*, 2022, pp. 69–87.

[23] G. Arya *et al.*, "Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training," *IEEE Access*, vol. 12, pp. 22359–22375, 2024, doi: 10.1109/ACCESS.2024.3361322.

[24] G. Boosting, "Analysis of the Best Social Media Platforms for Promotion Using Machine Learning and RFE Feature Selection : A Comparative Study of," vol. 10, no. 1, pp. 513–521, 2026.