Deep Learning Models For Youtube Sentiment Analysis: A Comparative Study Of Bert And Gru In Danantara Indonesia

M. Alfa Rizy [1], Ema Utami [1]
[1] Department of Informatics Engineering, AMIKOM University of Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received July 10, 2025
Revised July 17, 2025
Accepted Sept 26, 2025

Keywords:

Sentiment Analysis
Deep Learning
BERT
GRU
YouTube

ABSTRACT

This study aims to analyse public sentiment toward Danantara Indonesia through YouTube comments and compare the performance of two deep learning models, BERT and GRU. The methodology involved collecting and preprocessing a dataset of 1,065 comments, which revealed a significant class imbalance, with negative sentiments comprising the majority (60%), followed by positive (25%) and neutral (15%) sentiments. This data skew presented a key challenge for the modeling process. In the evaluation, the IndoBERT model demonstrated markedly superior performance, achieving a 100% accuracy score on the test set with perfect precision, recall, and F1-scores. In stark contrast, the GRU model yielded only 60% accuracy, exhibiting a strong bias towards the dominant negative class, likely due to the data imbalance. While the sentiment distribution suggests public distrust regarding Danantara's policies, BERT's flawless performance on this specific dataset, although impressive, also points to a potential risk of overfitting. Therefore, this study concludes that while BERT is a significantly more robust model for this task, its perfect score highlights the critical need for further evaluation. Future work should incorporate robust validation methods and data balancing strategies to verify the model's generalization capabilities across diverse social media contexts.

Corresponding Author:
M. Alfa Rizy,
AMIKOM University of Yogyakarta, Jl. Ring Road Utara Condongcatur, Yogyakarta, 55281, Indonesia
Email: alphariz@students.amikom.ac.id

1. INTRODUCTION

In the digital era, social media platforms have become key ecosystems for people to share opinions, experiences, and perceptions on various issues, including economic policies and the performance of strategic institutions. YouTube, as one of the largest social media platforms, serves not only as a means of entertainment but also as a venue for broad public discussion [1]. The various comments uploaded by users reflect spontaneous reactions to phenomena, including economic and investment policies undertaken by state institutions. In this research, Danantara Indonesia as a national sovereign wealth fund needs to understand how the public responds to the investment policies it manages [2]. As a strategic body aimed at optimizing government investments to drive national economic growth, Danantara Indonesia plays a vital role in building public trust, attracting investor interest, and supporting long-term economic stability [3].

Public trust in a sovereign wealth fund is a fundamental element in the successful management of state investment funds, especially for developing countries like Indonesia, where transparency and accountability remain primary concerns. However, along with increasing public engagement in the digital space, an institution's image and reputation are now significantly influenced by narratives developing on social media, both positive and negative. In practice, public sentiment recorded on social media can be an early indicator of public perception towards policies adopted by an institution. Previous studies show that opinions developing on social media can significantly impact investor confidence and a country's economic stability [4]. Therefore, understanding public sentiment towards Danantara Indonesia becomes crucial in assessing the effectiveness of investment policies and identifying potential reputational risks [5].

However, the main challenge in extracting information from YouTube comments is the large volume of data and variations in language style, compounded by the prevalence of complex linguistic phenomena such as subjectivity, sarcasm, irony, and other emotive expressions that can obscure user intent. To overcome these challenges, sentiment analysis based on artificial intelligence is increasingly used as a method for systematically understanding public opinion. Deep learning models such as Bidirectional Encoder Representations from Transformers (BERT) [6] and Gated Recurrent Unit (GRU) [7] offer more accurate approaches compared to conventional methods. BERT, with its Transformer architecture, is capable of understanding word context more deeply, while GRU, as a variant of Recurrent Neural Network (RNN), is more efficient in processing sequential text. This distinction in their architectural strengths—deep contextual understanding versus efficiency in sequential processing—makes a comparative study particularly insightful for practical applications.

While various studies have proven the effectiveness of deep learning-based sentiment analysis in diverse domains such as public service evaluation [8], government policy research [9], and economic trend prediction [10], a significant research gap persists in applying these methods to the highly specific and critical domain of sovereign wealth funds. Public sentiment analysis towards institutions like Danantara Indonesia—which operate at the intersection of high-stakes national finance and public trust—remains a largely unexplored area, especially within the context of Indonesian digital discourse. Therefore, this study aims to fill this gap by conducting a comparative analysis of BERT and GRU on YouTube comments related to Danantara Indonesia. By assessing the performance of these two models, this research is expected to provide deeper insights into public perception of state investment policies and identify the most effective model for this specific context. Furthermore, the results are intended to provide strategic recommendations for Danantara Indonesia to design more effective public communication, enhance transparency, and manage its institutional image, thereby strengthening its position as a credible and globally competitive sovereign wealth fund.

2. METHOD

2.1. Data Collection

The data used in this research consists of public comments in Bahasa Indonesia regarding Danantara Indonesia, sourced from the YouTube social media platform. Data collection was conducted over a three-month period, from February 2025 to April 2025.

The comment collection process targeted videos relevant to the topic of Danantara Indonesia, with video selection criteria focused on:

1. Content discussing policies, performance, or news related to Danantara Indonesia.
2. Video sources from trusted official news channels and channels belonging to influencers or experts known for discussing economic or public policy issues with a data-driven or scientific approach.

Based on these criteria, 14 YouTube videos were selected as data sources. Comments from these 14 videos were then collected using a scraping technique. The scraping process was carried out utilizing the “requests” and “pandas” library in Python, yielding a total of 31,675 raw comments.

The collected raw comments were then saved in CSV format for further preprocessing and data labeling stages. (Here you can add a sentence if there was a very basic initial cleaning process, for example: "In the initial phase, data cleaning was performed to remove identical duplicate comments." If not, you can proceed directly to the preprocessing stage in the next section).

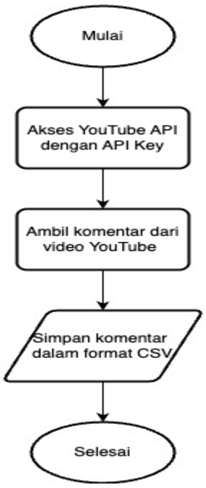

Mulai
Akses YouTube API dengan API Key
Ambil komentar dari video YouTube
Simpan komentar dalam format CSV
Selesai

Figure 1. Data Collection Step

2.2. Data Preprocessing

Prior to sentiment analysis, data preprocessing is an essential stage to clean and prepare the raw textual data extracted from YouTube comments, making it suitable for processing by the deep learning models. This stage aims to reduce noise, normalize the text, and structure the data into a format amenable to model training. The primary preprocessing steps performed in this study are detailed below:

1. **Case Folding:** All text was converted to lowercase. This step ensures uniformity across the dataset, preventing the models from treating the same word with different capitalization (e.g., "Danantara" and "danantara") as distinct entities.
2. **Text Cleaning (Noise Removal):** Several types of noise common in social media text were removed:
    a. **URLs:** All web links (HTTP/HTTPS) and short URLs were identified and eliminated.
    b. **Emojis:** Emojis were removed to simplify the textual content, as their interpretation can be ambiguous or require specialized handling beyond the scope of the primary sentiment classification task.
    c. **Special Characters and Punctuation:** Non-alphanumeric characters (e.g., !, ?, #, *, etc.), with the exception of those deemed relevant for sentiment expression if specifically handled, were removed. Excessive punctuation was also normalized.
    d. **Numbers:** Numerical digits were removed unless they were part of specific terms relevant to the sentiment context (which was generally not the case for this dataset).
3. **Tokenization:** Following the initial cleaning, the processed text of each comment was tokenized into individual words or sub-word units. This process breaks down the sentences into a sequence of tokens, which serve as the basic input for the subsequent natural language processing tasks.
4. **Stopword Removal:** Common Indonesian stopwords – words that occur frequently but carry little semantic weight for sentiment discrimination (e.g., "yang", "di", "dan", "adalah") – were removed. This was performed using [mention the stopword list or library used, e.g., "the Sastrawi library's stopword list," or "a curated list of Indonesian stopwords adapted for social media content"]. Removing stopwords helps in reducing the dimensionality of the data and allows the models to focus on more informative terms.
5. **Stemming:** To further normalize the vocabulary and group words with similar meanings, stemming was applied. The Sastrawi library, a popular stemmer for Bahasa Indonesia, was utilized to reduce words to their root or base form (e.g., "menganalisis," "dianalisis" to "analisis"). This process helps in consolidating the feature space and mitigating data sparsity.

2.3. Data Labelling

Following the completion of the preprocessing stage for the entire comment dataset, comprising 31,675 data entries, the next crucial step is data labeling. The objective of this stage is to categorize each comment into one of three predetermined sentiment classes: positive, neutral, or negative. This process generates the labels that will serve as the ground truth for training and evaluating the deep learning models in this study.

In this research, data labeling was performed automatically (auto-labeling). This approach was chosen due to the large volume of comment data, enabling a more efficient and rapid tagging process. Furthermore, auto-labeling aims to maintain consistency in label assignment and reduce potential subjectivity that can arise from manual labeling by multiple individuals. For this auto-labeling process, the study utilized a pre-trained IndoBERT-based sentiment classification model designed for analyzing Indonesian text. The specific model employed was mdhugol/indonesia-bert-sentiment-classification from the Hugging Face platform, which is capable of classifying text into three sentiment categories.

Each preprocessed comment was subsequently fed into this pre-trained model to assign a sentiment label. This auto-labeling process yielded a final distribution across the 31,675 comments as follows: 19,549 negative, 7,155 positive, and 4,971 neutral. This resulting distribution is visually depicted in Figure 2. The chart clearly illustrates a significant class imbalance, where the negative sentiment constitutes the vast majority of the dataset. This pronounced data skew is a critical characteristic of the dataset that will heavily influence the training and evaluation of the subsequent deep learning models.

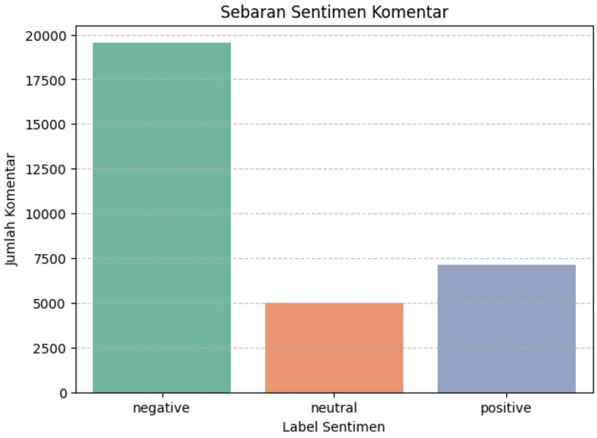


Figure 2. Sentiment Distribution Data

2.4. BERT (Bidirectional Encoder Representations from Transformers)

The first model evaluated in this comparative study is Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art language model renowned for its ability to capture deep bidirectional contextual relationships within text. The specific architecture of the BERT model as implemented for this sentiment classification task is illustrated in Figure 3.

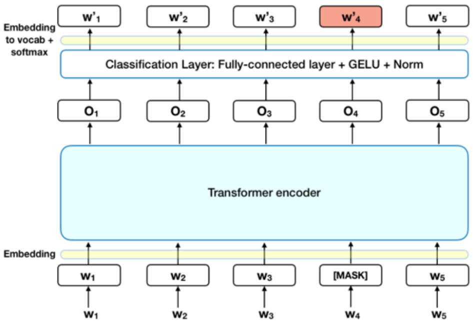

w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Figure 3. BERT Mask Language Modelling (MLM)

As depicted in the figure, the process begins by adding a special [CLS] token to the start of the input text sequence. This entire sequence is then converted into numerical vectors through an embedding layer. These embeddings are subsequently fed into the multi-layer Transformer encoder, which processes the sequence holistically to generate context-rich output representations for each token. For the final classification, the model uses the output vector corresponding to the [CLS] token, as this vector serves as an aggregate representation of the entire sequence's meaning. This single vector is then passed through a classification layer, which consists of a fully-connected neural network with a GELU activation function and layer normalization, ultimately producing the probability for each sentiment class (positive, neutral, or negative).

- Core Architecture and Pre-training: Illustrates the NSP task where the model learns to predict if one sentence logically follows another, provides context on how BERT learns sentence relationships during its initial training. However, for this sentiment analysis task, we primarily leverage a pre-trained IndoBERT model and fine-tune it.
- Input Representation for Sentiment Classification: For sentiment classification, input comments were tokenized using the IndoBERT tokenizer (typically WordPiece). Each tokenized input was formatted with special tokens, such as [CLS] at dissected the beginning (whose final hidden state is often used as the aggregate sequence representation for classification tasks) and [SEP] to mark the end of a sentence.
- Fine-tuning Architecture for Sentiment Classification: The pre-trained IndoBERT model was fine-tuned for the specific task of sentiment classification. This involved adding a classification layer on top of the IndoBERT base. Typically, the output representation from the [CLS] token of the final IndoBERT layer is fed into a dense (fully connected) layer with a softmax activation function to output probabilities for the three sentiment classes (positive, neutral, negative).
- Hyperparameters and Training Details:
    - Learning Rate:
    - Batch Size:
    - Number of Epochs:
    - Optimizer:
    - Max Sequence Length:

2.3. GRU (Gated Recurrent Unit)

The Gated Recurrent Unit (GRU), as proposed by Cho et al., is a type of Recurrent Neural Network (RNN) designed to overcome weaknesses of standard RNNs, particularly the vanishing gradient problem that often arises when processing long text sequences. GRU features a simpler gating mechanism compared to Long Short-Term Memory (LSTM) – consisting of an update gate and a reset gate. This simpler architecture makes GRU more computationally efficient while still being capable of capturing temporal dependencies and long-range patterns in text data effectively.

Figure 4, illustrating the workings of a GRU cell, effectively shows how this component within the neural network helps in retaining relevant information from sequences like text or time series.

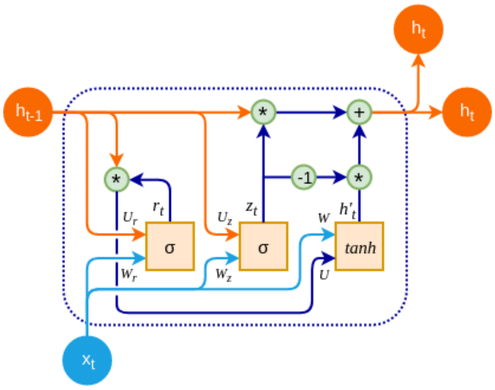


Figure 4. GRU Cell

2.4. Model Evaluation and Comparison

Following the implementation of the IndoBERT and GRU models for sentiment classification of YouTube comments regarding the subsidized LPG policy, an evaluation was conducted to assess the effectiveness of each model. This evaluation aimed to determine the more suitable deep learning model by considering aspects of accuracy, efficiency, and processing complexity. To measure the models' performance, several common evaluation metrics in text classification were employed, including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

Accuracy was used to measure how frequently the model made correct predictions across the entire test dataset. Precision measures how many of the positive sentiment predictions were correct relative to the total positive predictions made, whereas recall measures the model's ability to identify all actual positive instances. The F1-score was used as a measure of the balance between precision and recall, which is particularly useful in cases of an imbalanced distribution of positive and negative sentiment data. Furthermore, the confusion matrix was utilized to understand the classification errors made by the model, while the ROC-AUC score measures how effectively the model can distinguish between the positive and negative classes. To ensure the model was not overly dependent on specific patterns within the test data, the evaluation was conducted using a 10-fold cross-validation technique.

3. LITERATURE REVIEW

Research conducted by Kalanjati et al. highlights the impact of public sentiment on social media regarding the COVID-19 outbreak in Indonesia, which posed a challenge for policymakers in understanding public social behavior[11]. This study focused on the Twitter platform and found that public opinion towards government policies for managing the pandemic varied significantly.

Machine learning approaches in sentiment analysis have been applied in various studies concerning public opinion. Bhalerao et al.[12] as well as Valarmathi et al.[13] emphasized that machine learning techniques, particularly classification algorithms, are effective in categorizing public sentiment on social media. Research by Susanto et al.[14] also showed how classifier algorithms can be used to group public opinion into specific sentiment categories. In their study, methods based on Naïve Bayes and Support Vector Machine (SVM) demonstrated reasonably good performance in classifying positive, negative, and neutral sentiment in Indonesian-language text.

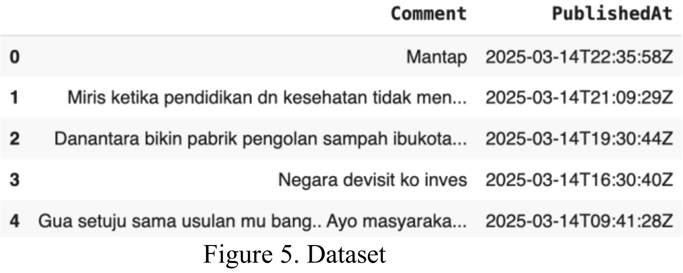
In line with advancements in sentiment analysis, Wan et al.[15] developed the ECR-BERT (Emotion-Cognitive Reasoning integrated BERT) model to improve the accuracy of sentiment analysis on Online Public Opinions regarding Emergency Events (OPOEs). This model integrates emotion modeling into BERT-based analysis, enabling sentiment analysis with higher accuracy and greater explainability in the context of emergency events. This study indicates that transformer-based models, such as BERT, possess an advantage in capturing the complexity of sentiment compared to conventional models that only consider lexical features.

On the other hand, models based on Recurrent Neural Networks (RNN), such as the Gated Recurrent Unit (GRU), have been proven to be a lighter and more efficient alternative in sentiment analysis tasks. Muhammet Sinan Başarslan and Fatih Kayaalp[16] proposed the MBi-GRUMCONV model, a deep learning approach that combines a Bidirectional GRU (Bi-GRU) and a Convolutional Neural Network (CNN) for sentiment analysis on the IMDB movie review dataset. By using six Bi-GRU layers and two CNN layers, this model was able to significantly increase sentiment prediction accuracy. Furthermore, the research also compared two-word representation methods, namely Skip-Gram and Continuous Bag of Words (CBOW) from Word2Vec, finding that the Skip-Gram method was superior in capturing complex context within the text. Experimental results showed that the MBi-GRUMCONV model achieved an accuracy of 95.34%, which surpasses the results of previous studies in the deep learning-based sentiment analysis literature.

4. RESULTS AND DISCUSSION

4.1. Data Collection

Data for this study were collected using a scraping technique from the comment section of a YouTube video. The collected data consist of comments from a single video discussing Danantara. The resulting dataset from the scraping process is presented in Figure 5 below.

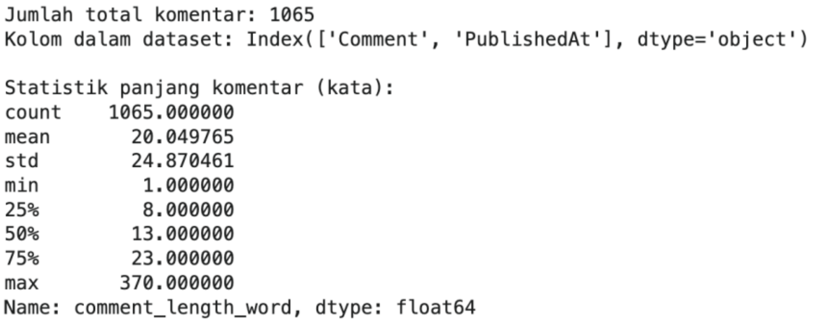


	Comment	PublishedAt
0	Mantap	2025-03-14T22:35:58Z
1	Miris ketika pendidikan dn kesehatan tidak men...	2025-03-14T21:09:29Z
2	Danantara bikin pabrik pengolan sampah ibukota...	2025-03-14T19:30:44Z
3	Negara devisit ko inves	2025-03-14T16:30:40Z
4	Gua setuju sama usulan mu bang.. Ayo masyaraka...	2025-03-14T09:41:28Z

Figure 5. Dataset

4.2. Exploratory Data Analysis (EDA)

In this study, YouTube comments were collected using web scraping techniques via the YouTube API. Following the data downloading and filtering process, a total of 1,065 comments were compiled, as detailed in the descriptive statistics in Figure 6.



```
Jumlah total komentar: 1065
Kolom dalam dataset: Index(['Comment', 'PublishedAt'], dtype='object')

Statistik panjang komentar (kata):
count    1065.000000
mean       20.049765
std        24.870461
min         1.000000
25%         8.000000
50%        13.000000
75%        23.000000
max       370.000000
Name: comment_length_word, dtype: float64
```

Figure 6. Descriptive Statistics

Figure 7 displays a Word Cloud generated from the dataset. The larger the size of a word in the Word Cloud, the higher its frequency in the dataset.

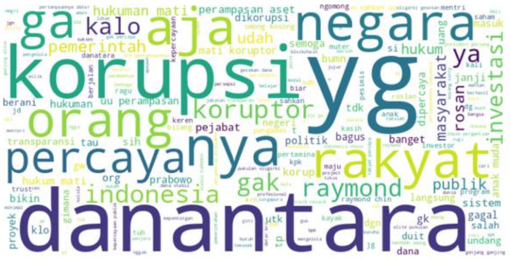


Figure 7. Dataset Word Cloud

Table 1 presents the 10 most frequently occurring words in the comments.

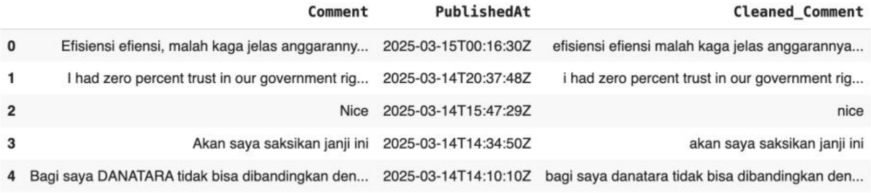
Tabel 1. Top 10 Most Frequent Words

Words	Frequency
yg	329
danantara	251
korupsi	224
koruptor	125
percaya	123
aja	114
raymond	113
indonesia	111
nya	110
rakyat	108

The word "danantara," referring to Danantara Indonesia, has a significantly high frequency (251 instances). This indicates that the discussion within the comments is indeed focused on this institution. The presence of words like "korupsi" (corruption, 224 instances) and "koruptor" (corruptor, 125 instances) shows that many comments address the issue of corruption in the context of Danantara Indonesia. This suggests the presence of negative sentiment or public distrust towards the related institution. The word "percaya" (trust/believe, 123 instances) appears frequently, which could reflect discussions related to the level of public trust or mistrust in the policies or officials involved. The name "Raymond" (113 instances) appears in the list of frequent words. This may indicate a specific public figure who is a subject of discussion in the comments. The appearance of "indonesia" (111 instances) and "rakyat" (the people, 108 instances) suggests that the comments extensively discuss matters of national interest and the relationship between policy and the populace.

4.3. Data Preprocessing

Before proceeding with further analysis, the comment data collected from YouTube via web scraping must be cleaned to be more structured and ready for processing by the deep learning models. This preprocessing stage involves applying a series of text transformations to eliminate irrelevant elements, which yields a clean, processable dataset as depicted in Figure 8.

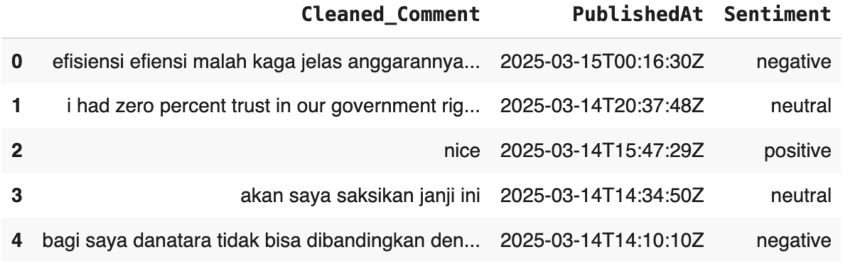


	Comment	PublishedAt	Cleaned_Comment
0	Efisiensi efiensi, malah kaga jelas anggaranny...	2025-03-15T00:16:30Z	efisiensi efiensi malah kaga jelas anggarannya...
1	I had zero percent trust in our government rig...	2025-03-14T20:37:48Z	i had zero percent trust in our government rig...
2	Nice	2025-03-14T15:47:29Z	nice
3	Akan saya saksikan janji ini	2025-03-14T14:34:50Z	akan saya saksikan janji ini
4	Bagi saya DANATARA tidak bisa dibandingkan den...	2025-03-14T14:10:10Z	bagi saya danatara tidak bisa dibandingkan den...

Figure 8. Preprocessed Dataset

4.4. Data Labeling

The data labeling process was conducted to classify the Indonesian-language text into three sentiment categories: positive, neutral, and negative. Figure 9 illustrates the outcome of this stage, where a new column titled 'Sentiment' was added to the dataset, reflecting the classification result for each comment. This labeled dataset serves as the foundation for subsequent analysis, including the performance comparison of the deep learning models (IndoBERT vs. GRU).



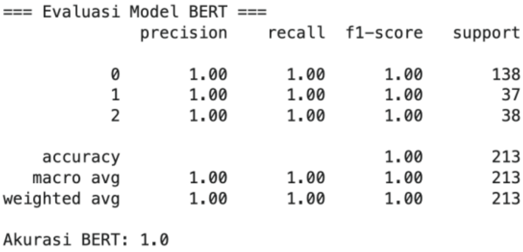
	Cleaned_Comment	PublishedAt	Sentiment
0	efisiensi efiensi malah kaga jelas anggarannya...	2025-03-15T00:16:30Z	negative
1	i had zero percent trust in our government rig...	2025-03-14T20:37:48Z	neutral
2	nice	2025-03-14T15:47:29Z	positive
3	akan saya saksikan janji ini	2025-03-14T14:34:50Z	neutral
4	bagi saya danatara tidak bisa dibandingkan den...	2025-03-14T14:10:10Z	negative

Figure 9. Labelled Dataset

4.5. Sentiment Classification

In this study, the sentiment classification process was performed using two deep learning models: IndoBERT and GRU (Gated Recurrent Unit). The IndoBERT model is a variant of Bidirectional Encoder Representations from Transformers (BERT) that has been specifically adapted for the Indonesian language. This model operates by learning the contextual relationships between words in a sentence bidirectionally, allowing it to capture deeper and more accurate meanings for sentiment classification.

Conversely, the GRU is a lighter recurrent neural network-based model compared to Long Short-Term Memory (LSTM). GRU is designed to process sequential data by retaining important information through a gated mechanism. In this research, the GRU model was trained using Word2Vec/FastText embeddings to efficiently recognize sentiment patterns within the YouTube comments.



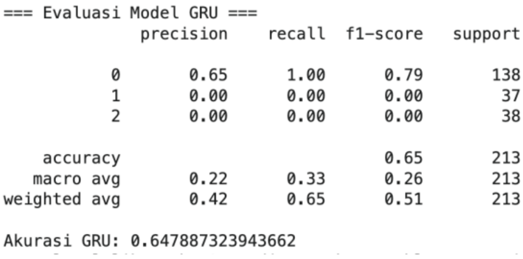
```
=== Evaluasi Model BERT ===
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       138
           1       1.00      1.00      1.00        37
           2       1.00      1.00      1.00        38

    accuracy                           1.00       213
   macro avg       1.00      1.00      1.00       213
weighted avg       1.00      1.00      1.00       213

Akurasi BERT: 1.0
```

Figure 10. Evaluation Results of the BERT Model



```
=== Evaluasi Model GRU ===
              precision    recall  f1-score   support

           0       0.65      1.00      0.79       138
           1       0.00      0.00      0.00        37
           2       0.00      0.00      0.00        38

    accuracy                           0.65       213
   macro avg       0.22      0.33      0.26       213
weighted avg       0.42      0.65      0.51       213

Akurasi GRU: 0.647887323943662
```

Figure 11. Evaluation Results of the GRU Model

Based on the evaluation results presented in Figure 10 and Figure 11, the IndoBERT model demonstrated significantly superior performance compared to the GRU model in the task of sentiment classification of YouTube comments concerning Danantara Indonesia. The IndoBERT model achieved a perfect accuracy of 100%, with precision, recall, and F1-scores of 1.00 for all sentiment classes (positive, neutral, and negative). These results indicate that IndoBERT was able to recognize sentiment patterns in the data exceptionally well, delivering entirely accurate predictions without any classification errors.

On the other hand, the GRU model struggled to classify sentiment in a balanced manner. Its accuracy only reached 60%, with highly inconsistent performance across sentiment classes. The GRU model showed strength in detecting the negative class with a recall of 100% but completely failed to classify comments into the positive and neutral classes, as indicated by precision, recall, and F1-scores of 0.00 for both. This suggests that the GRU model was heavily biased towards classifying comments into the dominant class, unable to properly distinguish other sentiment nuances.

Figure 12, presenting the confusion matrix, shows that BERT achieved perfect accuracy (100%) in classifying the YouTube comments' sentiment. There were no misclassifications whatsoever, as evidenced by all predictions falling on the main diagonal of the confusion matrix. This resulted in precision, recall, and F1-scores of 1.00 (100%) for all classes, indicating optimal performance across all evaluation aspects. However, although these results suggest the model performed exceptionally well, such perfect accuracy could be an indication of overfitting.

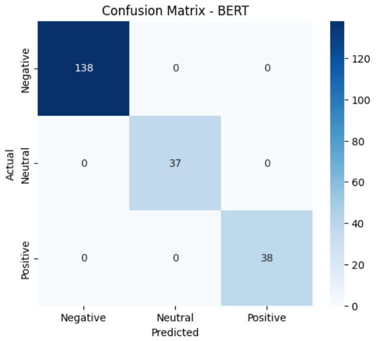


Figure 12. Confusion Matrix of the BERT Model

Conversely, as shown in Figure 13, the GRU exhibited extremely poor performance in sentiment classification, with an accuracy of only 60%. Consequently, the precision, recall, and F1-scores for the neutral and positive classes were 0.00, meaning the model failed to recognize these two categories. This indicates that the GRU model suffered from a very strong bias towards the negative class, causing nearly all comments to be categorized as negative.

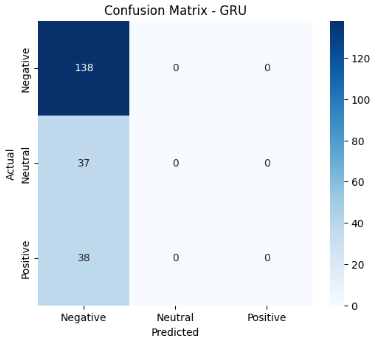


Figure 13. Confusion Matrix of the GRU Model

The IndoBERT model utilizes a transformer-based architecture that processes text bidirectionally. This allows the model to understand the meaning of a word in relation to the words that both precede and follow it within the same sentence. For instance, in the sentence "Saya tidak percaya dengan kebijakan ini" (I do not believe in this policy), the model processes the entire sentence simultaneously, enabling it to recognize the negative sentiment. In contrast, GRU is a recurrent neural network (RNN) that works sequentially, passing information from the previous word to the next. This approach can cause information from the beginning of a sentence to be lost by the time the end is reached, especially in longer texts. Consequently, GRU is more susceptible to losing critical context required for accurate sentiment classification.

IndoBERT employs WordPiece Tokenization, which better handles out-of-vocabulary words or words with complex morphology. Furthermore, the model has been pre-trained on massive Indonesian-language corpora, endowing it with a rich understanding of linguistic patterns, including slang, specific terminologies, and word variations common in YouTube comments. Meanwhile, the GRU in this study used Word2Vec or FastText embeddings. While capable of representing words as numerical vectors, these methods have limitations in understanding deeper semantic relationships. For example, Word2Vec primarily associates words based on their statistical proximity in a text, without grasping contextual meaning. As a result, the GRU model tends to be weaker in distinguishing sentiment in sentences with complex structures or ambiguous contexts.

IndoBERT's transformer-based architecture contains hundreds of millions of parameters and a self-attention mechanism that allows the model to capture inter-word relationships more accurately. The attention layers enable the model to assign higher weights to words that are more influential in determining a sentence's sentiment. Conversely, GRU has a lighter architecture with fewer parameters. While more computationally efficient, this architecture is less capable of capturing complex textual relationships. This makes GRU more prone to classification bias, often defaulting to classifying text into the dominant class (in this case, negative) without accurately identifying other sentiment nuances.

IndoBERT is a pre-trained model that has been trained on a vast amount of data before being fine-tuned for the sentiment analysis task. This pre-training provides a significant advantage in understanding broad language and sentiment patterns. Moreover, the BERT model can be easily adapted to specific data through fine-tuning, allowing it to acclimate to the linguistic characteristics of YouTube comments. On the other hand, the GRU in this study was trained from scratch (or with limited transfer learning using more restricted word embeddings). Consequently, the model is limited in its ability to recognize more complex patterns in the data, especially if the training dataset is relatively small.

5. CONCLUSION

The results show that BERT is significantly superior to GRU in this sentiment classification task, with the IndoBERT model achieving perfect accuracy (100%) and precision, recall, and F1-scores of 1.00 for all sentiment classes, a success attributed to its ability to understand context bidirectionally via its self-attention mechanism. Conversely, the GRU model only achieved 60% accuracy and completely failed to recognize neutral and positive sentiments—misclassifying all of them as negative—presumably due to architectural limitations and potential model bias. Therefore, it can be concluded that BERT is a more suitable model for sentiment analysis of Indonesian text on complex data like YouTube comments, although further evaluation through cross-validation is necessary to ensure generalization and prevent overfitting. As a final implication, this research highly recommends using transformer-based models like BERT for public policy sentiment analysis on social media and opens avenues for future studies exploring more advanced LLMs or improved fine-tuning methods on larger datasets.

REFERENCES

[1] G. R. Team, “YouTube Statistics 2025 [Users by Country + Demographics],” Official GMI Blog. Accessed: Mar. 09, 2025. [Online]. Available: https://www.globalmediainsight.com/blog/youtube-users-statistics/

[2] A. Gelb, S. Tordo, H. Halland, N. Arfaa, and G. Smith, *Sovereign Wealth Funds and Long-Term Development Finance: Risks and Opportunities*. in Policy Research Working Papers. The World Bank, 2014. doi: 10.1596/1813-9450-6776.

[3] Danantara, “Daya Anagata Nusantara - Danantara.” Accessed: Mar. 08, 2025. [Online]. Available: https://danantara.id

[4] R. An, “The Role of Digital Media in Shaping Public Relations: Developing Successful Online Communication Strategies for Enterprises,” *JADHUR*, vol. 3, no. 3, pp. 51–68, Sep. 2024, doi: 10.56868/jadhur.v3i3.246.

[5] B. AlBadani, R. Shi, and J. Dong, “A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM,” *ASI*, vol. 5, no. 1, p. 13, Jan. 2022, doi: 10.3390/asi5010013.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[7] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” 2014, *arXiv*. doi: 10.48550/ARXIV.1406.1078.

[8] A. Amalia, D. Gunawan, and K. Nasution, “Sentiment analysis of GO-JEK services quality using Multi-Label Classification,” *J. Phys.: Conf. Ser.*, vol. 1830, no. 1, Art. no. 1, Apr. 2021, doi: 10.1088/1742-6596/1830/1/012003.

[9] M. Y. Aldean, P. Paradise, and N. A. Setya Nugraha, “Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac),” *INISTA*, vol. 4, no. 2, Art. no. 2, Jun. 2022, doi: 10.20895/insta.v4i2.575.

[10] W. A. Degife and B.-S. Lin, “A Multi-Aspect Informed GRU: A Hybrid Model of Flight Fare Forecasting with Sentiment Analysis,” *Applied Sciences*, vol. 14, no. 10, p. 4221, May 2024, doi: 10.3390/app14104221.

[11] V. P. Kalanjati *et al.*, "Sentiment analysis of Indonesian tweets on COVID-19 and COVID-19 vaccinations," *F1000Res*, vol. 12, p. 1007, Apr. 2024, doi: 10.12688/f1000research.130610.4.

[12] A. A. Bhalerao, B. R. Naiknaware, R. R. Manza, and S. K. Bawiskar, "Sentiment Analysis on Covid-19 Vaccination Using Machine Learning Techniques," in *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*, vol. 105, S. Tamane, S. Ghosh, and S. Deshmukh, Eds., in Advances in Computer Science Research, vol. 105. , Dordrecht: Atlantis Press International BV, 2023, pp. 235–250. doi: 10.2991/978-94-6463-136-4_22.

[13] B. Valarmathi, N. S. Gupta, V. Karthick, T. Chellatamilan, K. Santhi, and D. Chalicheemala, "Sentiment Analysis of Covid-19 Twitter Data using Deep Learning Algorithm," *Procedia Computer Science*, vol. 235, pp. 3397–3407, 2024, doi: 10.1016/j.procs.2024.04.320.

[14] A. Susanto, M. A. Maula, I. U. W. Mulyono, and M. K. Sarker, "Sentiment Analysis on Indonesia Twitter Data Using Naïve Bayes and K-Means Method," *JAIS*, vol. 6, no. 1, pp. 40–45, May 2021, doi: 10.33633/jais.v6i1.4465.

[15] J. Wang, J. Du, Y. Shao, and A. Li, "Sentiment Analysis of Online Travel Reviews Based on Capsule Network and Sentiment Lexicon," *arXiv.org*, 2022, [Online]. Available: https://www.semanticscholar.org/paper/ce384aaa9bac2dadf565dd496fdccc8a06665c7d

[16] M. S. Başarslan and F. Kayaalp, "MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis," *J Cloud Comp*, vol. 12, no. 1, p. 5, Jan. 2023, doi: 10.1186/s13677-022-00386-3.